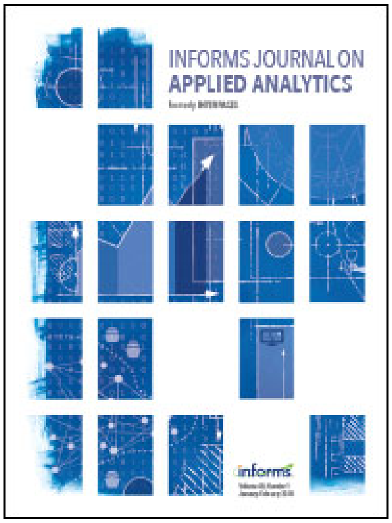


INFORMS Journal on Applied Analytics

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Bombardier Aftermarket Demand Forecast with Machine Learning

Pierre Dodin, Jingyi Xiao, Yossiri Adulyasak, Neda Etebari Alamdari, Lea Gauthier, Philippe Grangier, Paul Lemaitre, William L. Hamilton

informs
https://pubsonline.informs.org/journal/ijaa

INFORMS JOURNAL ON APPLIED ANALYTICS
Articles in Advance, pp. 1–21
ISSN 2644-0865 (print), ISSN 2644-0873 (online)

Bombardier Aftermarket Demand Forecast with Machine Learning

Pierre Dodin,[a] Jingyi Xiao,[b] Yossiri Adulyasak,[c,*] Neda Etebari Alamdari,[d] Lea Gauthier,[d] Philippe Grangier,[d] Paul Lemaitre,[d] William L. Hamilton[e]

[a] Bombardier, Saint-Laurent, Quebec H4R 1K2, Canada; [b] HEC Montréal, Montreal, Quebec H3T 2A7, Canada; [c] GERAD and Department of Logistics and Operations Management, HEC Montréal, Montreal, Quebec H3T 2A7, Canada; [d] IVADO Labs, Montreal, Quebec H2S 3J9, Canada; [e] Mila–Quebec AI Institute and School of Computer Science, McGill University, Montreal, Quebec H3A 2A7, Canada

*Corresponding author

Contact: pierre.dodin@aero.bombardier.com (PD); jingyi.xiao@hec.ca (JX); yossiri.adulyasak@hec.ca, https://orcid.org/0000-0002-6996-0742 (YA); neda.etebarialamdari@ivadolabs.com (NEA); lea.gauthier@ivadolabs.com (LG); philippe.grangier@ivadolabs.com (PG); paul.lemaitre@ivadolabs.com (PL); wlh@cs.mcgill.ca (WLH)

Received: November 2, 2021
Revised: June 2, 2022; December 18, 2022
Accepted: January 18, 2023
Published Online in Articles in Advance: May 2, 2023

https://doi.org/10.1287/inte.2023.1164

Abstract. Intermittent demand patterns are commonly present in business aircraft spare parts supply chains. Because of the infrequent arrivals and large variations in demand, aircraft aftermarket demand is difficult to forecast, which often leads to shortages or overstocking of spare parts. In this paper, we present the development and implementation of an advanced analytics framework at Bombardier Aerospace, which is carried out by the Bombardier inventory planning team and IVADO Labs to improve the aftermarket demand forecasting process. This integrated predictive analytics pipeline leverages machine-learning (ML) models and traditional time series models in a single framework in a systematic fashion. We also make use of a tree-based machine-learning method with a large set of input features to estimate two components of intermittent demand, namely demand sizes and interdemand intervals. Through the ML models, we incorporate different features, including those derived from flight data. Outputs of different forecasting models are combined using an ensemble technique that enhances the robustness and accuracy of the forecasts for different groups of aftermarket spare parts categorized by demand patterns. The validation results show an improvement in forecast accuracy of approximately 7% and in unbiased forecast of 5%. The ML-based Bombardier Aftermarket forecasting system has been successfully deployed and used to forecast the aftermarket demand at Bombardier of more than 1 billion Canadian dollars on a regular basis.

History: This paper was refereed.

Keywords: aftermarket spare parts • business aircraft • intermittent demand forecasting • machine learning

1. Introduction

The *Bombardier Aftermarket* business unit provides various services and products (e.g., spare parts for in-service aircraft) to support business aircraft maintenance. This unit manages a very complex supply chain system that operates the spare-part inventory distribution in more than 15 worldwide locations to ensure service quality and customer satisfaction in a timely manner. Among Bombardier's aftermarket inventory, 90% of the spare parts exhibit intermittent demand patterns. Providing more accurate spare part forecasts has a direct impact on business operations by mitigating the risk of inventory obsolescence (Porras and Dekker 2008, Teunter et al. 2011) or inventory shortage (Syntetos and Boylan 2005, Kourentzes 2013). This aftermarket demand forecasting process is a critical activity in aftermarket supply and inventory planning at Bombardier, which allows the company to efficiently manage the procurement and inventory of spare parts and increase its ability to maintain decent service-level agreements and customer satisfaction.

The previous aftermarket demand forecasting process at Bombardier mainly relied on internal forecasting software from a third-party software provider. This forecasting tool makes use of the exponential smoothing model of Croston (1972), which is a simple univariate time-series model, and other proprietary algorithms to produce demand forecasts. Often, forecasts produced by this model are inaccurate, requiring the domain experts at Bombardier to check and adjust the forecasts regularly. This manual adjustment process is cumbersome and unsystematic. Moreover, the knowledge and experience of domain experts vary quite significantly, which affects the consistency and accuracy of forecasting performance. It is also a very time-consuming process, especially when dealing with a large number of parts, which is the case at Bombardier Aftermarket.

This motivated Bombardier to develop an integrated and robust machine learning (ML) solution to improve the forecasting performance of the demand for spare parts in more than 15 worldwide locations. The development of this ML-based aftermarket demand forecast system was a collaboration between Bombardier Aftermarket and IVADO Labs that took place between September 2020 and February 2021. ML-based models are particularly useful in this context because we can incorporate multiple input features, such as flight activity data. In addition, one can train an ML model across a group of items rather than the historical time series of a single item, as in the case of traditional univariate time-series models, such as Croston's method.

This paper presents the development of two ML-based demand forecasting frameworks, the rotable ML (RML) framework and the intermittent ML (IML) framework, which make use of various features, including flight data for demand prediction. More specifically, RML is developed to deal with nonintermittent and mildly intermittent demands, which are typical among rotable parts—that is, parts that can be repaired and replaced. This framework comprises the gradient boosting model, called the light gradient boosting machine (LGBM) (Ke et al. 2017) and Elastic Net regression (Zou and Hastie 2005), which is a regularized linear regression model. We choose these two different types of ML-based models based on the fact that their different representations can capture different aspects of demand patterns. Meanwhile, IML is developed to deal with intermittent and highly intermittent items, which are more challenging. IML comprises three models, namely, Croston's model, LGBM, and the newly proposed variant of Croston's model in which we develop two ML models to predict the size and frequency of demand, respectively. We refer to Croston's model and its ML variant model as the decomposed IML in this paper, and they are only used to predict a subset of intermittent items. We also leverage an efficient ensemble technique in both the RML and IML framework to combine the forecasts produced by different models. More details on the demand classification scheme and solution methodology are provided in Section 3.2.

The contributions of our work to the practice and implementations of analytics can be described as follows. First, we leverage a forecast ensemble approach that combines traditional time series and ML models for a real-world aftermarket demand forecasting problem. With ML models, we are able to include multiple relevant external variables such as flight activity features into the models, which cannot be directly achieved by using traditional time series methods. Second, we propose the use of a tree-based regression method (LGBM) with a large set of input features to estimate the two components of an intermittent demand model (i.e., nonzero demand size and interdemand interval) of Croston's model as part of the ensemble process. Third, we present the overall pipeline and feature engineering process that has been successfully deployed to handle more than 1 billion dollars of spare parts on a regular basis. Fourth, through a cross-validation process, we numerically validate the performance of the proposed analytics framework and show that this ML-based forecasting system significantly improves the demand forecast accuracy. Finally, our work has been fully deployed into a production environment. This case study demonstrates how the artificial intelligence (AI)/analytics solution can facilitate daily spare parts forecast production and how it contributes to improvements in spare parts demand planning and inventory management of the Bombardier Aftermarket business.

2. Literature Review

There is a large body of literature on demand pattern categorization schemes and demand forecasting models, which are suitable for different categories of items based on their characteristics (Williams 1984, Syntetos and Boylan 2001, Johnston et al. 2003, Syntetos et al. 2005, Boylan et al. 2008, Kaya et al. 2020). Williams (1984) categorized demand patterns using a lumpiness measure based on the variance partition technique. Syntetos et al. (2005) suggested a revised classification scheme based on the average demand interval (*ADI*) and the squared coefficient of variation of nonzero demand sizes (CV^2).

Conventionally, there are two primary categories of demand forecast: (i) smooth and (ii) intermittent (or nonsmooth) demand. Intermittent demand, which is integer-valued and nonnegative, appears when the time between two consecutive demand events is relatively long and inconsistent. This demand pattern frequently occurs in business aircraft spare parts supply chains because their occasional uses and flight patterns are difficult to predict (Amirkolaii et al. 2017). In addition, because of infrequent demand arrivals and large variations, conventional time series forecasting methods, such as simple exponential smoothing (SES) and moving average (MA), often fail to produce reliable predictions when dealing with intermittent demand (Croston 1972, Gamberini et al. 2010). To address this type of demand, Croston (1972) first proposed an enhanced exponential smoothing method to explicitly capture demand intermittency, which was later refined by Rao (1973). More specifically, this method deals with the irregular occurrences and high variations of nonzero demands in an intermittent demand pattern by separately forecasting (i) the size of nonzero (positive) demand $\hat{z}_t$ at time period t and (ii) the estimated interval between two consecutive periods with nonzero demands (interdemand interval) $\hat{p}_t$ at time period t. Rather than applying SES to the original demand time series, Croston's method estimates $\hat{z}_t$ and $\hat{p}_t$ separately using the smoothing functions

$\hat{z}_t = \hat{z}_{t-1} + \alpha_z(z_{t-1} - \hat{z}_{t-1})$ and $\hat{p}_t = \hat{p}_{t-1} + \alpha_p(p_{t-1} - \hat{p}_{t-1})$, where z_t and p_t are the actual demand size and the number of periods from the period when the previous positive demand occurs until period t, respectively; α_z and α_p are smoothing parameters of the model. The final forecast is $\hat{y}_t = \hat{z}_t/\hat{p}_t$.

Croston's method has motivated a number of research works in intermittent demand forecasting (Schultz 1987, Willemain et al. 1994, Johnston and Boylan 1996, Strijbosch et al. 2000, Willemain et al. 2004, Syntetos and Boylan 2006). Although Croston's method improved the forecasting accuracy by smoothing the demand and time interval separately, it has some inherent issues in terms of modeling assumptions and estimation bias (Willemain et al. 1994). For instance, Croston (1972) assumed that demand sizes and time intervals were mutually independent and that time intervals between two consecutive demands as well as demand sizes were subsequently independent. Willemain et al. (1994) challenged the assumption of mutual independence because they found significant correlation between these two estimators in real-world data. Later, rather than relying on the successive demand independence assumption, Willemain et al. (2004) developed a model that incorporates autocorrelation assumption. Syntetos and Boylan (2001) demonstrated that Croston's method is biased in the final forecast ratio $\hat{z}_t/\hat{p}_t$. The method itself has a tendency to overforecast intermittent demand (such as spare parts), and the magnitude of the forecasting bias is positively correlated with the value of the smoothing parameter α_p. To resolve the issue of bias in Croston's method, Syntetos and Boylan (2005) applied a damping factor, $1 - \alpha_p/2$, directly to the final Croston's forecast ratio—that is, $\hat{y}_t = \left(1 - \frac{\alpha_p}{2}\right)\frac{\hat{z}_t}{\hat{p}_t}$, where α_p is equivalent to the smoothing parameter of interdemand interval estimation. This modified version is commonly referred to as Syntetos-Boylan approximation (SBA) (Petropoulos and Kourentzes 2015). Other research works have also proposed various adjustments to the final forecast ratio (Levén and Segerstedt 2004, Teunter et al. 2011), but multiple studies have reported the superior performance of the adjustment method presented by Syntetos and Boylan (2005) (Petropoulos and Kourentzes 2015). There are also research works that propose variants of Croston's method or the development of new methods to improve the accuracy of forecasting intermittent demand, such as bootstrapping-based methods (Willemain et al. 2004, Snyder et al. 2012) and neural networks (Amin-Naseri and Tabar 2008, Gutierrez et al. 2008, Kourentzes 2013).

Petropoulos and Kourentzes (2015) sought to improve intermittent demand forecasting by exploring the efficiency of forecast combinations. They combined Croston's method, SBA, and other simple forecasting methods (e.g., MA) and showed that combining multiple forecasts of intermittent demand from different models could improve forecasting accuracy and robustness. The combinations produce a performance superior to that of single-forecast methods and are runtime efficient without intensive model selection or a time-consuming hyperparameter tuning process.

As different forecasting models are created based on different assumptions and input features, combining (ensembling) forecasts from different models often leads to more reliable and better forecast accuracy in practice (Bates and Granger 1969, Bajari et al. 2015, Hyndman and Athanasopoulos 2018). Jose and Winkler (2008) demonstrates that aggregating predictions from several forecast models generally reduces sensitivity to estimation errors. The combinations of time-series forecasts from multiple models can take the form of a simple set of weights, such as equal weights (Makridakis and Winkler 1983, Clemen 1989, Jose and Winkler 2008), or a more complex set, such as weights derived from Akaike's information criterion (AIC) (Kolassa 2011). Petropoulos and Kourentzes (2015) present a review of commonly used forecast combination methods for intermittent demand forecast and empirically validate their performances. The results show that the combined forecasts are generally superior to forecasts generated by single methods. In addition, they point out that such ensemble methods eliminate the need for extensive model selection and hyperparameter search, which helps to simplify the overall forecasting pipeline and process.

There are some notable studies that leverage machine learning–based methods for spare-part demand forecasting. Gutierrez et al. (2008) make use of a neural network model comprising three layers with a sigmoid activation function and two input variables: most recent demand and the number of periods between the last two nonzero demands. Kourentzes (2013) extends the work of Gutierrez et al. (2008) by leveraging the same input features and three-layer neural network models. However, they make use of a tanh activation function and two neural network models to predict the demand rate and demand interval separately. Recently, Amirkolaii et al. (2017) have explored the use of neural network models with 16 input features based on time-series components (such as recent means, standard deviations, number of periods between the last two positive demands, most recent demand, and other time series–related components) and price as an additional input variable. They empirically measured the forecasting accuracy against traditional time series methods such as moving average, simple exponential smoothing, and Croston's model using demand data of 30 selected items (SKUs). Our study differs from these aforementioned works in several aspects. First, in addition to time series–related features, we extract features based on flight activity data as additional input variables for the ML models. Second, we leverage tree-based models, which allow interpretability that is not possible under neural network models (Zhang et al. 2021). Such tree-based models are also used to predict demand size and demand interval, which are combined in the same way as in the Croston's method. Third, rather than relying on a single model, we make use of an ensemble technique to combine the outputs from multiple models to create more robust forecasts in a systematic fashion. Finally, the effectiveness of our framework, which is fully deployed, is demonstrated through a large-scale implementation at Bombardier for spare-parts demand prediction of approximately 70,000 items. Indeed, the distinct contribution of our paper in the spirit of applied analytics and applications lies in the fact that we incorporate domain knowledge and customize the feature engineering process to derive domain-specific input variables for aircraft aftermarket demand forecast. Likewise, the adaptations of a tree-based modeling approach and an ensemble method allow us to successfully develop a reliable predictive analytics pipeline that yields superior and robust performance compared with traditional time series methods and the legacy software.

3. Overview of the Forecasting System at Bombardier

3.1. Bombardier Aftermarket Spare-Parts Landscape

Bombardier Aftermarket manages approximately 70,000 spare parts and classifies the parts, by convention, into two primary categories: *expendable* parts (comprises approximately 68,000 parts) and *rotable* parts (comprises approximately 2,000 parts). Expendable parts comprise newly manufactured parts and are typically disposable after use, whereas rotable parts are refurbished spare parts that receive maintenance after entering into Bombardier's parts repair pool to extend their use duration.

The spare parts transactions are spread over three geographical regions: (i) AMERICA, which covers the North and South American regions; (ii) EMEA, which comprises the Europe, Middle East, and Africa regions; and (iii) APAC, which is the Asia-Pacific region. Under the region level, there are hubs, service centers, and depots. Figure 1 illustrates an overview of Bombardier Aftermarket's parts landscape over three regional hubs (Chicago/Frankfurt/Singapore), eight service centers, and six depots.

3.2. Demand Classification and Model Selection

As illustrated in Section 1, there are generally two categories of demand pattern: intermittent and nonintermittent. In this case study, we calculate ADI and CV^2, which are the key measures used to capture the characteristics of demand (Syntetos et al. 2005) for both rotable and expendable parts to identify their degree of intermittency. This calculation serves as a criterion to select which models to implement on rotable and expendable parts, respectively. At the same time, we use ADI and CV^2 as input features to capture statistics related to demand intermittency. Although one can also use these values to directly determine clusters of the part and allow part to change from one cluster to another when these measures for a specific part change, we do not consider such an option because it is important in this project to ensure that part clusters must remain static over time and aligned with the business practice at Bombardier. In addition, the data models in the pipeline are also maintained separately for rotable and expendable parts.

The average demand interval is the summation of time intervals between two consecutive nonzero demands divided by the total number of time intervals—that is, $ADI = \frac{\sum_{t=1}^{N} p_t}{N}$, where p_t is the interdemand interval of nonzero demands at period t, and N is the total number of periods. The squared coefficient of variation measures the variation of nonzero demand sizes and is calculated as the square of the standard deviation of nonzero demand sizes divided by the mean of nonzero demands—that is, $CV^2 =$

Figure 1. (Color online) Bombardier Aftermarket's Spare Parts Consist of Two Part Categories (Expendable and Rotable Parts): These Parts are Stored and Distributed from Regional Hubs, Service Centers, and Depots

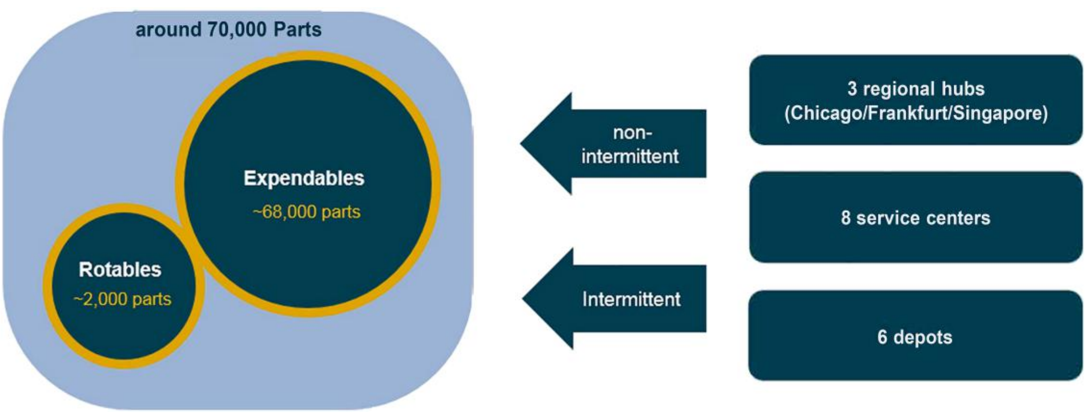
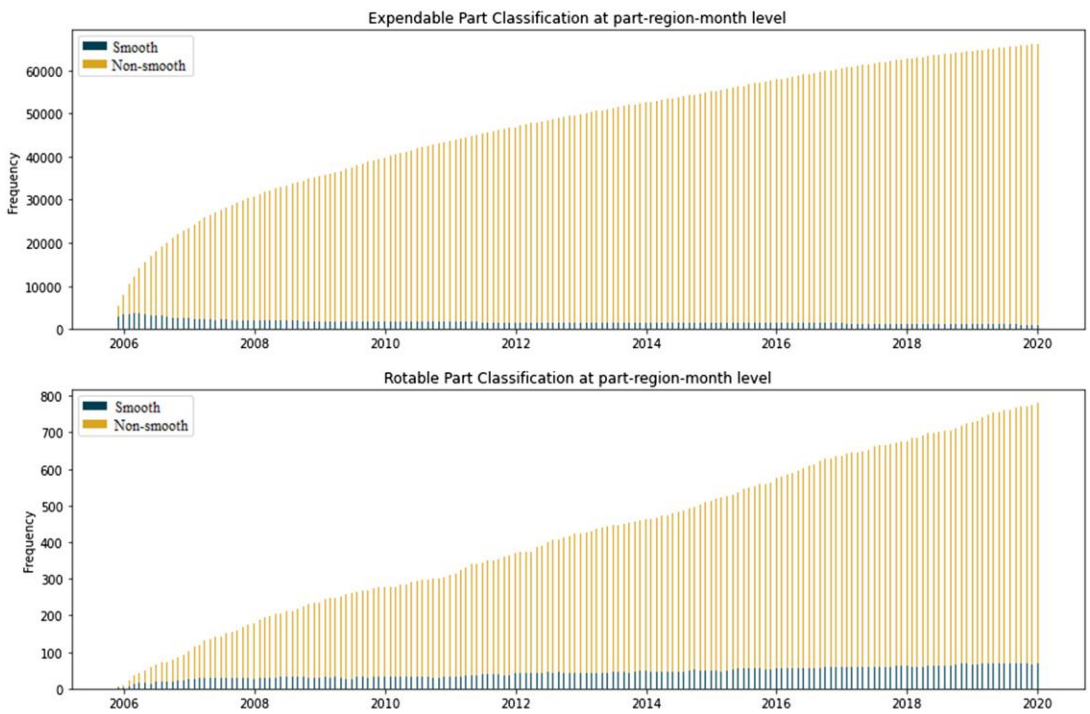


Figure 2. (Color online) Proportions of Smooth and Nonsmooth Demands in Expendable and Rotable Parts Categories



$\left[\frac{\sqrt{\frac{\sum_{t=1}^{N}(z_t-\overline{z})^2}{N}}}{\overline{z}}\right]^2$ and $\overline{z}=\frac{\sum_{t=1}^{N} z_t}{N}$, where z_t is the nonzero demand size at period t and $\overline{z}$ is the expected value of nonzero demands.

Syntetos et al. (2005) further characterize demand into four categories: *smooth*, *erratic*, *intermittent*, and *lumpy*. They identified an item as smooth when $CV^2 \leq 0.49$ and $ADI \leq 1.32$, erratic when $CV^2 > 0.49$ and $ADI \leq 1.32$, intermittent when $CV^2 \leq 0.49$ and $ADI > 1.32$, and lumpy when $CV^2 > 0.49$ and $ADI > 1.32$. Figure 2 depicts the demand pattern distribution in both parts at the part-region-month level. As illustrated in the figure, both part categories exhibit a primarily nonsmooth demand pattern. The nonsmooth items account for more than 90% of the expendable parts, whereas rotable parts include a higher proportion of items with the smooth demand pattern. Nevertheless, they still exhibit a considerable level of intermittence wherein the majority of items in this group have the values of ADI and CV^2, which exceed the thresholds recommended by Syntetos et al. (2005) for smooth demand.

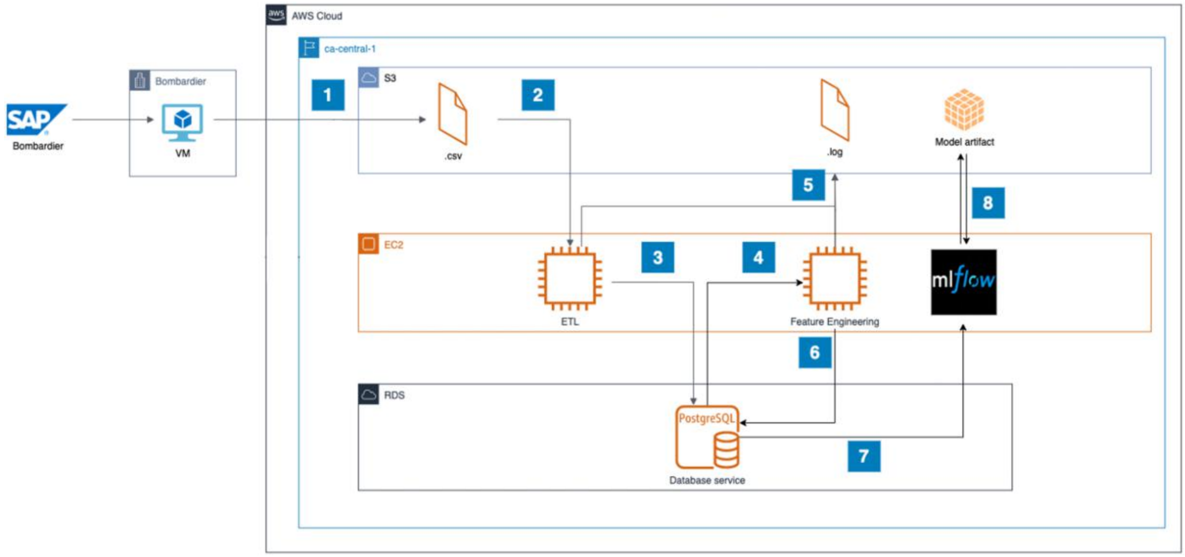
Based on the existing business practices and requirements at Bombardier Aftermarket as described earlier, we implemented an ML-based framework dedicated to forecasting rotable parts and another ML-based framework that handles expendable parts. In each framework, we develop and combine multiple forecasting models through an ensemble method. We build this ML pipeline architecture on Amazon Web Services (AWS), and Figure 3 demonstrates the ML architecture workflow. In the next section, we discuss the forecasting models used for different groups of items.

3.3. Overview of Input Features and Model Outputs

Unlike traditional time series models (e.g., Croston's model), which are trained using historical time series demand data, the ML-based models can be trained across multiple parts and products to learn underlying patterns that can be observed in the same group of parts. In practice, for such ML-based models to be effective, one would need to include a set of input features that help represent the characteristics and behavior of the demand of a specific part. To this end, we make use of various input features related to recent demand observations and their statistics, spatial and temporal information, and other features such as those related to the use of the parts described later in this section to capture heterogeneity among parts. Similar feature engineering processes and their effectiveness have also been demonstrated in other works that leverage ML models (Ferreira et al. 2016, Makridakis et al. 2022).

We perform the feature engineering process in two steps. First, by conducting a series of workshops with the Bombardier Aftermarket team and performing data quality checks, we list available data and use the experts' insights to create a list of potential input features for

Figure 3. (Color online) Architecture of the Predictive Analytics Pipeline Implemented in Amazon Web Services



Notes. In the first step (process 1), the pipeline extracts raw inputs from the Bombardier database system, which comprises multiple data sources ranging from sales orders, service, customer information, scheduled and unscheduled maintenance, to aircraft flight activities. This extraction process combines data sources from multiple servers (process 2) to create data copies pending for extract, transform, load (ETL) data preprocessing. We store the data copies on AWS S3, which are then transferred to the ETL data processing pipeline (process 3). This ETL pipeline performs basic data preprocessing, such as date time conversion, data filtering, and feature engineering processes (processes 4 and 5). Then, we store the preprocessed data in a dedicated database (process 6) and subsequently feed the data into ML models in the ML flow platform (process 7). Finally, we connect the outputs of the predictive models to Bombardier's main enterprise system (process 8).

the machine-learning models. Together with the project team, we then prepare different sets of input features created based on various hyperparameters and feature transformations. In the second step, we set up a pipeline and conduct preliminary tests based on different input features to evaluate the models' performances and determine the most appropriate set of input features for the machine-learning models. We design the feature engineering process to be flexible, and it can be easily reconfigured by the users if needed.

We construct different input features to capture the characteristics of demand patterns. More specifically, we derive the following key features associated with the demands from the preprocessed data. First, we construct *lag demands* (i.e., demand sizes) of past m months, where the value of m depends on different groups of items. For instance, in the most recently deployed pipeline, we use one- to four-month lag of historical demands for rotable parts ($m = 4$) and one- to six-month lag of demands for expendable parts ($m = 6$). This is because expendable parts have a strong intermittent demand pattern, so they require the lag feature to go further into the historical horizon. Second, we compute *rolling demand features*, which measure statistics of demands within an n-month historical rolling window, as shown in Table 1. The rolling window can vary from two to n months, where the value of n can be set independently for each group of items.

In addition to the statistics of the demands, we also calculate the long-term average demand from the earliest demand in the data until the most recent date. Similar to the SES model, we also use the most recent forecast of the current month (lag-1) as an input feature. The feature engineering process in the pipeline performs a selection of the parameters m and n, which improves the overall prediction accuracy (we provide a list and a summary of the parameter optimization process in Appendix C). Third, to capture seasonal fluctuation effects in time series, we also construct the *seasonal fluctuation feature*, which transforms month indices into numerical values that preserve the continuity of the time horizon. In particular, the month index would result in a discontinuity when there is a transition between two different years (i.e., between December and January of the subsequent year). To this end, we use the sine- and cosine-encoding techniques to transform the month index into its corresponding sine and cosine values, which are the commonly used methods for capturing seasonal components of time series (Pollock 1993). In addition to these features, we also include ADI and CV^2 as other features to capture demand intermittency.

Besides the features related to spare parts demand, the input features also comprise a region and a month index as categorical features and other features related to an overall flight activity and part use. Because a part is installed on multiple aircraft, we aggregate flight activity data of all the aircraft in which this part was installed. Appendix A describes the feature construction procedure of flight activity features. Other relevant features

Table 1. Descriptions and Examples of Input Features Used in ML Models (Elastic Net and LGBM)

Type	Feature	Potential value
Demand lag	Lag values of spare-parts demand of the past m months	Capture the correlation between past and future demand
Demand statistics	Rolling mean, median, lag-1 forecast, standard deviation, and skewness demand features over varying windows of 2 to n months	Capture behaviors of previous demand information
Intermittent characteristics	• *ADI* during the last 12 months • CV^2 during the last 12 months	Capture intermittency and volatility of spare parts demand
Seasonality fluctuation	Seasonality components using sine and cosine transformations	Preserve information between two adjacent months regardless of year (e.g., December and January)
Time and region index	• Month index as a categorical variable • Region index as a categorical variable	Capture idiosyncratic behaviors of different months and regions
Features derived from flight activity data	Three features derived from flight hours and landing cycles associated with each part. The details of feature engineering for these variables are included in Appendix A.	Classify each part by different categories to capture the trends, behavior, and intermittent aspects of flight hours of the aircrafts that install the part
Others	• Number of unique aircrafts with this part • Part age	Provide supplemental information about parts and the business jets

include the number of aircraft in which this part was installed, the average part age. In particular, the average part age is calculated from the ages of all the aircraft with the part installed up until the month associated with this data point. In the experiments, we constructed these features from the available data for a period of 60 months from January 2015 to December 2019. The key demand features, other additional features, and their potential values are presented in Table 1.

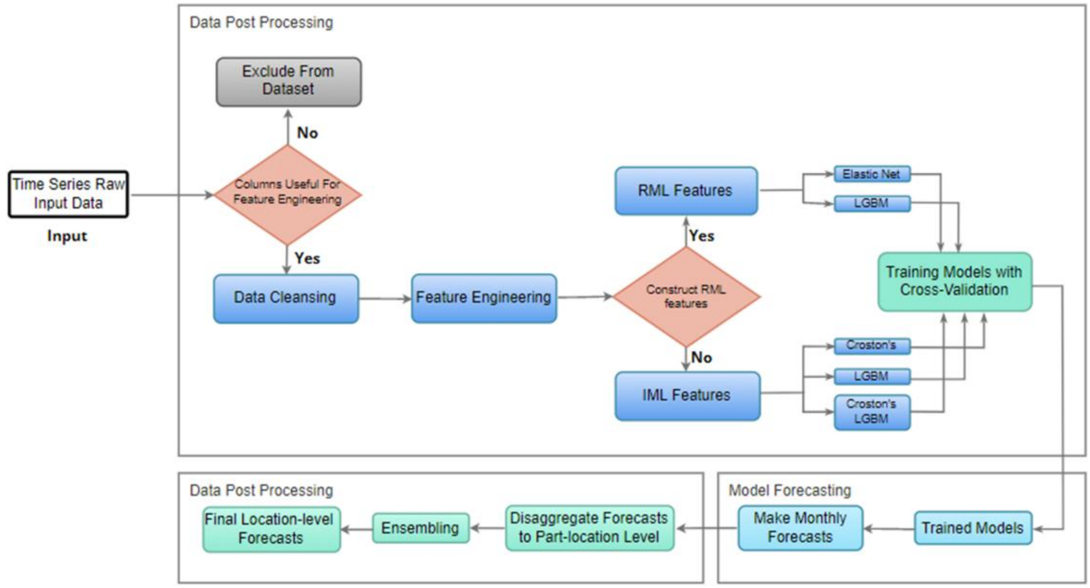
In addition, we highlight that the overall predictive performance of the models measured by the error metrics in the cross-validation process is the main objective of the pipeline. Because our task focuses on prediction (as opposed to causation), multicollinearity, which undermines statistical significance of an independent variable, is not a major issue (Allison 1999). This process of feature engineering and selection has been built into the pipeline, and the demand planners can perform this exercise periodically to evaluate and choose a set of features that result in an overall better prediction accuracy in the cross-validation process. Once we create the features, we combine them together to produce final input features (these input features can vary in different models based on data quality and availability). Nevertheless, we design the ML pipeline in such a way that the set of input features can be easily modified and changed in the pipelines.

In accordance with the existing forecasting process of Bombardier Aftermarket, the model outputs are demand forecasts at the part-region-month level (e.g., the predicted demand of part #111 in the EMEA region in June 2022 equals 15 units). Appendix B provides an illustrative example of input features and outputs. The motivation of making forecasts at the part-region level is both model oriented and business oriented. From the modeling perspective, the team avoids directly forecasting at the part-location level because demands at the location level can be extremely sparse. Forecasting at the part-region-month level increases model robustness and consistency. From the business perspective, in practice, if some spare parts are not available in a location, maintenance can be assigned or transferred to an alternative service location within a given region where all the required parts are available. In addition, Bombardier Aftermarket can also transfer spare parts between locations, especially in the same region if needed. To determine the part requirements at the location level for the inventory planning process, the region-level forecasts for each part number produced by the pipeline are distributed to each of the locations in the same region using predetermined disaggregation coefficients wherein each coefficient represents the percentage of monthly demand forecast of a given part assigned to each location in that region.

4. Experimental Design

In this section, we describe our modeling methodology, model architecture, and performance evaluation. There are two predictive frameworks, RML and IML, which provide the demand forecasts for rotable and expendable parts, respectively. These two frameworks differ in terms of the models and features used for demand forecasting. In both the RML and IML frameworks, we leverage different ML models that incorporate multivariate input features and time-series data to produce different forecasts, which are combined using an ensemble method. Generally speaking, the models used in the RML framework are more suitable for relatively smooth demand, whereas the models used in the IML framework are more suitable for intermittent demand. We

Figure 4. (Color online) Overall Predictive Analytic Process and Models Used in the Pipeline



provide an illustration of the models and predictive analytics process from a modeling perspective in Figure 4. We then describe the models and cross-validation process in the subsequent sections.

4.1. Predictive Models Used in the RML Framework for Rotable Parts

This section describes two ML models, LGBM and Elastic Net, which provide demand forecasts for the rotable parts. One of the advantages of the ML-based models as opposed to exponential smoothing methods (such as Croston's method) is the fact that multiple input features can also be used in the predictive model as predictors in addition to time-series components (Ferreira et al. 2016, Makridakis et al. 2022). Moreover, a tree-based regression model (such as LGBM) is a nonlinear model, which can potentially provide better predictive performance compared with models that assume specific structures such as traditional linear regression or support vector regression. The LGBM machine-learning library (Microsoft Corporation 2022) is also built with more than 10 different objective (loss) functions such as L1 loss, L2 loss, quantile loss, and Tweedie function, which has been proven to be effective in predicting slow-moving demand (Makridakis et al. 2022). Januschowski et al. (2021) also discuss similar advantages of the tree-based models used for demand forecasting. We then use an ensemble technique to combine the outputs of the LGBM and Elastic Net to produce the forecasts for the rotable parts. We refer to this approach as RML.

4.1.1. Elastic Net Regression. Elastic Net is a linear regression with L1 (lasso) and L2 (ridge) regularizations. Zou and Hastie (2005) proposed it as a unified regularization and feature-selection technique that takes advantage of the two commonly used regularization functions. In our project, because we employ an ensemble method that uses the forecasts produced by different models from different types, Elastic Net, which is based on linear regression, can capture the linear relationships among the predictors. The use of Elastic Net has proven to be effective in the cross-validation because it reduces the underforecasting issue in rotable parts that could occur when we use the nonlinear model (LGBM).

4.1.2. LGBM for RML. LGBM is a decision tree–based algorithm under a gradient boosting framework, proposed by Ke et al. (2017). LGBM can be used in classification and regression tasks. Compared with other tree-based boosting algorithms, one unique aspect of the LGBM is that it generates the tree by splitting the leaf that would lead to the maximum loss reduction, whereas other algorithms generally grow the tree by level. This growing algorithm thus contributes to faster training iterations and lower memory use, which allows LGBM to efficiently handle large-sized data. Recent studies and ML competitions have demonstrated the advantage and efficiency of LGBM (Makridakis et al. 2022). We provide more details of this algorithm and its hyperparameter optimization in Appendix C.

In rotable parts forecasting, LGBM can capture complex and nonlinear relationships among different features of demand and flight activities. Although the incorporation of a comprehensive feature set potentially helps to improve the forecast accuracy, LGBM tends to underforecast the rotable spare parts demand in our experiments. Therefore, we leverage the LGBM model alongside the Elastic Net model in an ensemble process to obtain more stable and robust results. We provide details of the ensemble method in Section 4.3.

4.2. Predictive Models Used in the IML Framework for Expendable Parts

Expendables generally exhibit intermittent-dominating patterns, and only approximately 10% of expendable parts exhibit smooth demand patterns. Because these items are highly intermittent, we leverage three different forecasting models in the IML framework for expendable parts. To this end, we implement an LGBM model similar to the one used in the RML framework and two decomposition models that are suitable for highly intermittent items—namely, Croston's and LGBM Croston-like models. These decomposition models essentially estimate two demand components, nonzero demands and interdemand intervals, separately. We provide the descriptions of the models employed in the IML framework in this section.

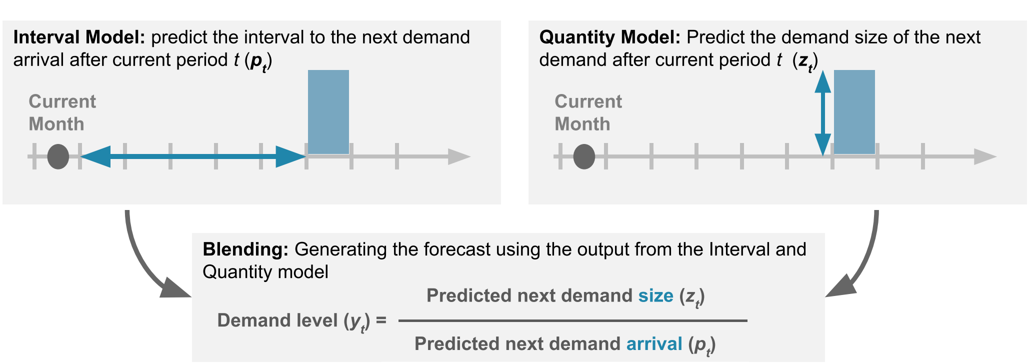
4.2.1. LGBM Model for IML. We construct this LGBM model in a similar way as the LGBM model for rotable parts except that we modify its input features to capture the intermittent behaviors. More specifically, there could potentially be many zero demands in the historical data, and the time horizon associated with the demand features must be expanded to increase the model's robustness against demand fluctuations. We extend lag demand features to the previous $m = 6$ months and rolling demand features to the previous $n = 12$ months. Additionally, because of limited data availability of expendable parts associated with each aircraft, we do not include features related to parts usage and flight activities in the LGBM for expendable parts. Nevertheless, we can later directly incorporate these features in the pipeline once sufficient data are available. The rest of the input features are consistent with the LGBM for RML.

4.2.2. Modeling Intermittent Demand. Prior to describing the two decomposition models, we first describe how these models are set up to capture intermittent demand patterns. Conventional time series methods that are commonly used for smooth demand often lead to poor results when dealing with irregular and sporadic demand of intermittent parts (Gamberini et al. 2010, Kaya et al. 2020). Therefore, when forecasting expendable parts, we leverage the idea of Croston decomposition and separately model the original time series with two components: the nonzero demands sizes (z_t) and the interdemand intervals (p_t). Karthikeswaren et al. (2021) present several classical and ML-based models that leverage similar approaches. In our case study, in addition to applying the original Croston's SES model, we also develop an ML-based decomposition model based on LGBM to directly estimate z_t and p_t. Based on these estimations, we calculate the final forecast for a given period by dividing $\hat{z}_t$ and $\hat{p}_t$: $\hat{y}_t = \frac{\hat{z}_t}{\hat{p}_t}$. We illustrate the targets (outputs) produced by the decomposition models in Figure 5.

As in Croston's method, both $\hat{z}_t$ and $\hat{p}_t$ are only updated at each time t when $\hat{z}_t$ differentiates from zero. Otherwise, $\hat{y}_t$ is not calculated, and the forecasted values remain unchanged from previous time periods $\hat{z}_{t-1}$ and $\hat{p}_{t-1}$.

4.2.3. SES Croston's Model. In the classical SES Croston's method, we decompose the time series of an item into nonzero demands sizes (z_t) and interdemand intervals (p_t). Both components are estimated by SES, wherein one submodel predicts the quantity of the next nonzero demand and another predicts the time interval before the next nonzero demand. Each component has a smoothing parameter (α_z and α_p). Although it is a common practice that both $\hat{z}_t$ and $\hat{p}_t$ share the same smoothing parameters, Schultz (1987), Snyder (2002), and Teunter et al. (2010) considered using different parameters for smoothing each estimation separately, which can improve the overall forecasting performance. Therefore, we also optimize the smoothing parameters of $\hat{z}_t$ and $\hat{p}_t$ separately by performing a grid search method. Apart from

Figure 5. (Color online) Two Intermittent Demand Components: Interdemand Intervals and Nonzero Demand Sizes



smoothing parameter optimization, we also follow the bias correction method proposed by Syntetos and Boylan (2001, 2005) to optimize the final demand rates $\hat{z}_t/\hat{p}_t$ using a damping factor $1-\alpha_p/2$, and we select the damping factor from a predefined range that results in the least forecasting biases. This method is referred to as Syntetos-Boylan approximation in our experiments.

4.2.4. LGBM Croston-Like Decomposition Model. This model is an ML-based variant of SES Croston's model in which the demand size $\hat{z}_t$ and interval $\hat{p}_t$ of period t are estimated by the LGBM algorithm. To this end, we implement two LGBM models with the identical format of input features but different outputs (dependent variables), demand size $\hat{z}_t$, and interval $\hat{p}_t$. To better capture the demand intermittency, rather than using the original lag demand features as in the LGBM for IML, we use two sets of demand features that are associated with $\hat{z}_t$ and interval $\hat{p}_t$. The first set of lag demand features is similar to the original lag demand features, but we only consider the positive demands. We first remove zero demands from the demand vector of an item. Then, we construct the lag demands of the past $m = 6$ periods (only the periods with positive demands are considered). The second set of lag demand features includes the lag demand intervals wherein we use the interval length of each of the two positive consecutive demands of the past $m = 6$ periods. In addition to these demand features, as in the LGBM model for IML, we also make use of the rolling features with $n = 12$, cyclical features, and time and region categorical features. Finally, we also apply the same bias correction method used in the SES Croston's model to improve the predictive performance of this LGBM decomposition model.

Even deep learning models, in particular, the recurrent neural network models (Hewamalage et al. 2021), which are more complex, can potentially be considered. This type of model is not considered by the project team because of two main important reasons. First, the training process of deep-learning models is highly computationally demanding, and the use of GPUs either locally or on the cloud platform in the training process is often necessary. Nevertheless, such an investment is not part of the scope in this project. Second, unlike tree-based models, which directly allow interpretability and insights, deep neural network models are not generally interpretable (Zhang et al. 2021). This aspect is important because the demand-planning team needs to regularly justify the forecasts in their demand-planning process.

4.3. Aftermarket Forecast Ensemble

In this work, we leverage an ensemble technique to combine the outputs of multiple forecasting models in both the RML and IML frameworks. To this end, we use the ensemble technique of the `opera` (Online Prediction by ExpeRt Aggregation) package developed by Gaillard and Goude (2020). In this approach, rather than using a predetermined fixed set of weights, `opera` combines forecasts from multiple learning algorithms using the weights that are derived from their past time-series prediction performances (Gaillard and Goude 2020). Multiple recent studies have demonstrated the effectiveness of the `opera` package (Cerqueira et al. 2019, Goehry et al. 2019).

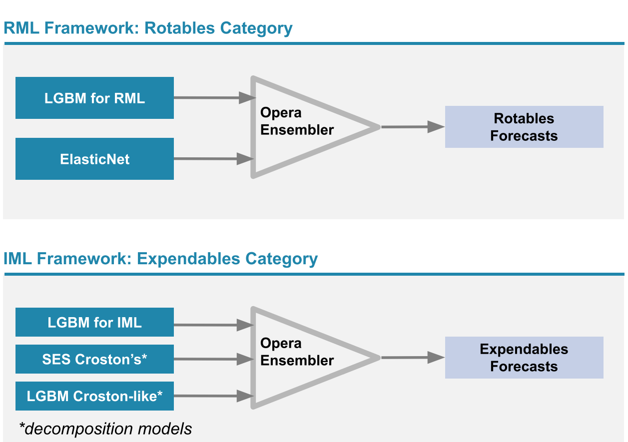
To determine our combined forecasts, denote by $\hat{y}_{i,t}$ the predicted demand for period t obtained from forecasting model i and by the aggregation weight $w_{i,t}$ associated with the prediction made by forecasting model i for time period t, and the aggregate (ensemble) prediction $\hat{y}_t^{ens}$ at time t can be computed as $\hat{y}_t^{ens} = \sum_{i=1}^{I} w_{i,t}\hat{y}_{i,t}$, where $w_{i,t}$ used to combine the forecasts for period t are optimized in `opera` package using the historical performance of previous forecasts $\hat{y}_{i,1}, \ldots, \hat{y}_{i,t-1}$ of each model i. We use the loss function based on squared errors of the historical forecasts to optimize these weights.

Figure 6 presents the forecasting flows and the ensemble processes of the RML and IML frameworks.

4.4. Performance Evaluation

We select two main performance metrics to measure the model errors and bias, namely, root-mean-squared scaled error (RMSSE) and normalized forecasting bias (NFB). RMSSE is a forecasting measure introduced by Hyndman and Koehler (2006) as a suitable performance measure for demand data that are generally close to zero. Recent literature (Niu 2020) and an international forecasting competition hosted on Kaggle (Makridakis et al. 2022), a large-scale data science platform, commonly adopts this measure. NFB, on the other hand, is an important measure adopted by the Bomdardier inventory management department that considers the cases of underforecasts and overforecasts separately to measure the implications of the forecasting performances on inventory planning. More particularly, NFB measures the cumulative bias normalized by the

Figure 6. (Color online) Ensemble Processes in RML and IML Frameworks



standard deviation of the errors produced by a reference model (i.e., the legacy system in our case). To determine whether part-region-level forecasting is considered acceptable, we compare NFB with a predefined threshold ϵ. Over a given horizon h, we define a part-region forecast as *unbiased* if $|\text{NFB}| \leq \epsilon$, *underforecasted* if $\text{NFB} > \epsilon$, and *overforecasted* if $\text{NFB} < -\epsilon$. We set this threshold of the normalized bias to be 3.0 in both predictive ML frameworks. Meanwhile, we implement k-fold time series cross-validation (Bergmeir et al. 2018) to evaluate performance of the models. We describe more details of the two main performance metrics, RMSSE and NFB, in Appendix D.

4.4.1. Aftermarket Forecast Cross-Validation Process.

k-Fold time series cross-validation is a validation method commonly used for tuning hyperparameters and evaluating time series predictions produced by different predictive models. This method is particularly useful for pipelines that make use of ML models to overcome overfitting (Bergmeir et al. 2018).

We implement cross-validation in both RML and IML pipelines. The training fold starts in September 2015, whereas the cross-validation folds are evenly sliced over the forecasting time interval from May 2018 to December 2019 (i.e., over the total period of 20 months), and each fold comprises four months. Thus, we obtain $k = 5$ folds in this validation process. This data split results in approximately 0.2 million entries (data points) per part-region-month in each CV fold.

5. Empirical Validation and Business Implications

We first empirically validate the performance of the ML-based RML and IML forecasting pipelines based on the two key performance measures. This empirical validation serves two main purposes. First, prior to full deployment, we must demonstrate that the forecast accuracy obtained by the ML-based pipelines is acceptable and more reliable than the forecasts produced by the legacy software. Second, Aftermarket must carefully evaluate the practicality and usability of the ML-based pipelines, particularly the data ETL pipeline, feature engineering, hyperparameter tuning, model training, and cross-validation. Once Aftermarket has performed the empirical validation, the executive sponsoring team of this project would review and approve the developed pipelines for deployment. After this process, the project team still continued to monitor the results and assessed the business values that this project elicits.

In this section, we first provide a summary of the results obtained in the empirical validation process prior to deployment using the data from May 2018 to December 2019 as the evaluation set. Then, we discuss the business implications and values based on the results and feedback received from the users after the pipelines have been fully deployed. As in the rest of the paper, we only provide the results and details that are not subject to confidentiality restrictions and have been reviewed by the company.

5.1. Performance Comparisons with Benchmarks and Individual Models

We report the prediction accuracy of individual models, including legacy software and SES Croston's model and Syntetos-Boylan approximation (SBA), two commonly adopted SES models for intermittent demand, as the benchmark approaches, in Table 2. The accuracy results are based on the two key important metrics used by the company, RMSSE and percentage of normalized biases within acceptable threshold (NFB), as described in the previous section. For rotable parts, both the LGBM and RML ensemble approaches outperform the legacy software, Croston's, and SBA models in terms of the prediction accuracy measured by RMSSE, whereas the Elastic Net regression and RML ensemble approaches yield

Table 2. Accuracy Results of Individual and Ensemble Models Based on RMSSE and Percentage NFB for Rotable and Expendable Parts

		RMSSE		NFB		
Part class	Model	Average	DIFF vs. legacy	Unbiased	Over	Under
Rotable parts (RML)	Legacy software	0.55	—	72.4%	20.4%	7.2%
	SES Croston	0.61	+10.9%	67.7%	23.0%	9.3%
	Syntetos-Boylan approximation	0.58	+5.5%	69.4%	18.6%	12.0%
	LGBM for RML	**0.49**	−10.9%	64.9%	7.2%	27.9%
	Elastic Net	0.62	+12.7%	76.8%	18.7%	4.5%
	RML ensemble framework	0.51	−7.3%	**77.5%**	10.3%	12.2%
Expendable parts (IML)	Legacy software	0.55	—	69.2%	24.4%	6.4%
	SES Croston	0.59	+7.3%	67.1%	24.1%	8.8%
	Syntetos-Boylan approximation	0.56	+1.8%	68.6%	19.1%	12.3%
	LGBM for IML	0.52	−5.5%	72.8%	15.5%	11.7%
	LGBM Croston-like	0.57	+3.6%	68.9%	24.7%	6.4%
	IML ensemble framework	**0.49**	−10.9%	**74.2%**	16.5%	9.3%

Note. Best results are indicated in bold.

superior results in terms of NFB. Even though the LGBM model provides the best RMSSE, this model generally underforecasts the demand (with 27.9% underforecasted demand below the threshold), whereas only 4.7% of the forecasts produced by Elastic Net are considered underforecasted. On the other hand, the RML framework, which combines multiple forecasting models, could yield almost the same level of RMSSE as LGBM (i.e., the RMSSE of the ensemble method is 2% higher than that of LGBM), whereas the percentage of normalized biases within acceptable threshold (NFB) is much higher (77.5% versus 64.9%). For expendable parts, the LGBM model has the lowest RMSSE among individual models, whereas the IML ensemble approach yields the best results in terms of RMSSE and NFB. The LGBM Croston-like model performed slightly better than the classical Croston's model but slightly worse than SBA. At the same time, the LGBM Croston-like model has the lowest percentage of underforecasted demands. Thus, by incorporating the LGBM Croston-like model in the ensemble process, it helps decrease the percentage of underforecasted demands in the IML ensemble, whereas the overall accuracy (measured by RMSSE) is minimized. These results clearly demonstrate the value of the ensemble technique, which helps significantly improve the performance and robustness of the predictions by combining the forecasts generated by multiple models.

Compared with the historical results produced by the legacy tool, the RML framework produced forecasts with an overall relative improvement of 7.3% on RMSSE (decreasing from 0.55 to 0.51) and an overall relative improvement of 7.0% on unbiased NFB (increasing from 72.4% to 77.5%), whereas the relative improvements achieved by the IML ensemble were 11.0% (decreasing from 0.55 to 0.49) and 7.2% (increasing from 69.2% to 74.2%) on RMSSE and unbiased NFB, respectively. These results were statistically significant at the confidence level set by the company.

Another significant advantage the team gained from this improved forecasting performance is a significant reduction in overforecasts. The ML-based frameworks reduced the overforecasts that exceed the acceptable threshold (3.0) set by the company by 49.5% for RML and by 32.4% for IML, translating into a potential inventory decrease. Although the percentage of underforecasts under the ML-based frameworks have increased, the margins are substantially lower, and the results remain within acceptable ranges based on the validation with the planning team. This is because slight or moderate increases in underforecasts do not necessarily lead to shortages because Bombardier also makes use of safety stocks in their inventory planning process to hedge against this risk.

In Table 3, we further analyze the effectiveness of different approaches and break down the performances of individual models by demand pattern—namely, smooth, lumpy, intermittent, and erratic—which is classified using

Table 3. RMSSE and NFB of Rotable and Expendable Items Under Different Demand Patterns

		RMSSE				NFB											
						Smooth			Lumpy			Intermittent			Erratic		
Part class	Model	Smooth	Lumpy	Intermittent	Erratic	Unbiased	Over	Under	Unbiased	Over	Under	Unbiased	Over	Under	Unbiased	Over	Under
Rotable parts (RML)	Legacy software	0.62	0.59	0.54	0.65	69.3%	24.0%	6.7%	71.0%	22.5%	6.5%	73.0%	19.8%	7.2%	62.1%	30.6%	7.3%
	SES Croston	0.66	0.70	0.53	0.75	70.0%	20.9%	9.1%	58.0%	30.4%	11.6%	67.9%	23.0%	9.1%	64.2%	25.8%	10.0%
	Syntetos-Boylan approx.	0.64	0.67	0.52	0.73	71.4%	14.4%	14.2%	60.1%	25.8%	14.1%	70.0%	18.9%	11.1%	65.3%	19.5%	15.2%
	LGBM for RML	**0.59**	**0.47**	**0.45**	0.59	77.5%	10.2%	12.3%	66.2%	6.3%	27.5%	63.4%	5.0%	31.6%	**77.8%**	11.4%	10.8%
	Elastic Net	0.71	0.61	0.63	0.67	**77.8%**	10.4%	11.8%	75.3%	20.2%	4.5%	76.6%	20.7%	2.7%	77.7%	9.7%	12.6%
	RML ensemble framework	**0.59**	0.53	0.48	**0.58**	74.9%	13.8%	11.3%	**79.2%**	10.0%	10.8%	**77.8%**	10.0%	12.2%	73.8%	15.4%	10.8%
Expendable parts (IML)	Legacy software	0.66	0.55	0.55	0.64	68.5%	24.0%	7.5%	65.3%	30.2%	4.5%	73.6%	17.6%	8.8%	65.0%	30.2%	4.8%
	SES Croston	0.65	0.55	0.55	0.63	72.8%	19.9%	7.3%	65.4%	27.5%	7.1%	70.9%	17.8%	11.3%	70.6%	22.7%	6.7%
	Syntetos-Boylan approx.	**0.61**	0.51	0.51	**0.58**	73.6%	13.6%	12.8%	67.0%	24.5%	8.5%	71.8%	15.2%	13.0%	71.9%	17.8%	10.3%
	LGBM for IML	0.64	0.48	0.52	0.60	72.7%	17.2%	10.1%	70.8%	12.8%	16.4%	75.0%	7.7%	17.3%	70.6%	18.4%	11.0%
	LGBM Croston-like	0.68	0.55	0.57	0.65	72.6%	19.5%	7.9%	65.9%	29.7%	4.4%	74.4%	18.6%	7.0%	**73.1%**	22.9%	4.0%
	IML ensemble framework	0.63	**0.47**	**0.47**	0.59	**74.3%**	16.2%	9.5%	**72.1%**	20.4%	7.5%	**77.1%**	11.2%	11.7%	72.3%	20.0%	7.7%

Notes. Best results are indicated in bold. approx., approximation.

the *ADI* and CV^2 thresholds proposed by Syntetos et al. (2005). For rotable parts, the results show that the machine-learning model (LGBM) and the ensemble model are the best-performing ones. The LGBM model generally performs well across all the four demand patterns in terms of RMSSE, whereas the performance of the RML ensemble model is slightly better than the LGBM model for erratic demand. However, the RML ensemble method results in lower prediction biases (i.e., higher NFBs) than the LGBM model for lumpy and intermittent items (these two groups comprise items with high average demand intervals, i.e., $ADI > 1.32$). This clearly demonstrates the effectiveness of the RML ensemble method, which leverages both the tree-based ML and time series models, to produce robust predictive performance. Among the models based on traditional time series methods, the legacy software produces the best results except for the case of intermittent demand, in which the SBA model provides superior results among such models. For expendable parts, the SBA model outperforms other methods for smooth and erratic demand ($ADI \leq 1.32$). Among the individual models, the SBA model performs well on three demand classes wherein the LGBM model is superior to the SBA when dealing with lumpy demand. Although the LGBM Croston-like model shows worse RMSSE than other individual models in the case of smooth and erratic demand, it results in high unbiased percentages under lumpy, intermittent, and erratic demand. We can also observe the benefit of the ensemble method, wherein it yields the best RMSSE in the case of lumpy and intermittent demand and provides robust predictions (measured by NFB) across all four demand patterns. The results of the legacy software, however, are inferior to other time-series methods for the case of expendable parts.

Because a tree-based model like LGBM allows interpretabilty and insights, we also extract feature importance results from the LGBM models for RML and IML and report the results in Appendix E. One notable observation is that, besides time series–related components, features extracted from the flight data are among the top contributing features, which demonstrate the benefit of incorporating such features into a machine-learning model for business aircraft spare parts demand prediction.

In addition to the numerical results of the forecast accuracy, the team also rigorously monitored and assessed both the IML and RML pipelines in terms of scalability and usability. Because of the significant volume of data involved in the forecasting process, the legacy forecasting software required approximately 30 hours to process and create the forecasts. The overall processing time in the legacy software is long because it requires an intensive data preparation process, and a large number of queries were required. Modification to that data pipeline is also not possible because it has been set up by the software provider. Meanwhile, the average runtime of newly developed forecasting pipelines is approximately eight hours, which include six hours for the ETL and prediction pipeline and two hours in the forecast ensemble process. Thus, this new system reduces runtime by more than 73%.

5.2. Business Implications after Deployment

A significant improvement in forecast accuracy would have a positive impact on spare parts inventory planning. With improved forecasting performance, Bombardier Aftermarket would lower its risk of inventory obsolescence and inventory shortage. This system is currently deployed and used regularly to handle Aftermarket's inventory. The Bombardier Aftermarket team can quickly obtain forecasts, make necessary adjustments, and use these forecasts to communicate and create fulfillment orders in a much more timely fashion. This helps the team to enhance its inventory planning and supply chain operations, as well as finance and business budgeting process. We also collected qualitative feedback from the planning team with respect to the indirect positive outcomes and implications of the project deliverables and summarized them as follows:

1. **Increase inventory savings:** The improved forecasts result in more accurate part distribution plans in the three operating regions. Therefore, this tool helps to reduce costs from holding superfluous inventory, such as storage, management, and overtime labor costs. In addition, the amounts of transshipments among different operating regions are reduced. According to Bombardier Aftermarket, the inventory turnover has improved more than 4% after the new forecasting system was fully deployed.

2. **Improve the production planning process and flexibility of spare parts:** Bombardier Aftermarket forecasts serve as a guideline for spare parts manufacturing and maintenance lines to prepare for the contingent parts demand. The levels of contingent parts are continuously monitored against the predicted demand, and alerts are created to notify the manufacturing team to increase the production of spare parts to meet the contingent demand. The ML-based pipelines also improved the scalability of the forecast process significantly with approximately 70% runtime reduction and approximate three times reduction in personnel hours in the demand planning process. Thus, improvements in spare parts forecasts (both the predicted demand sizes and demand arrivals) and overall process allow the manufacturing team to promptly and properly plan for and proactively react to the situation at hand.

3. **Improve customer satisfaction with spare parts availability:** Between January and July 2021, after Bombardier Aftermarket put the forecasting system into actual use, it kept actively monitoring the forecast performance, and the overall forecast accuracy improved by approximately 5.3% compared with the legacy software, with an overall unbiased forecast of 75.8%, which is close

to the accuracy levels achieved in the empirical validation phase. This improved accuracy helps increase part availability at Bombardier, and the maintenance is able to accept the maintenance requests and deliver more timely business aircraft maintenance service to its customers. Such improvements in forecast accuracy and inventory increase the fill rate (the percentage of customer demand that can be met by immediate stock) and the number of customers who can be served on time. This in turn increases customer satisfaction and loyalty.

4. **Mitigate financial and foreign currency risks:** Bombardier Aftermarket, which provides services to 15 different worldwide locations, has more than 1 billion dollars of total spare parts value in its supply chain network. When faced with an inventory shortage, Bombardier Aftermarket must transfer the required inventory to the destinations. Because parts must be exported and imported in this process, high-value and bulky parts can be very expensive to move. In addition, such transactions involve an inventory value transfer across different countries and an immediate currency conversion, which can be unfavorable for the company compared with planned exchange contracts. Therefore, the reduction in inventory transfers across certain countries helps to reduce the risk of unscheduled currency exchange, and the more accurate forecasts can help prepare the company to proactively allocate funds under favorable currency exchange contracts.

6. Conclusion

Bombardier Aftermarket, in collaboration with IVADO Labs, has jointly developed an automated ML-based demand forecasting system to predict aftermarket spare parts demand for Bombardier business aircraft. To this end, we built two ML-based solution frameworks, namely RML and IML, to generate forecasts for rotable and expendable parts, respectively. This forecasting system leverages traditional time series methods for intermittent demand and modern machine-learning algorithms to generate forecasts of aftermarket spare parts. To fully take advantage of the ML models, we created multiple sets of features related to spare parts demands, seasonality, and time and region categorical features, as well as features derived from flight activity and maintenance data. The forecasts generated by different models are combined through an adaptive forecast ensemble method, which assigns weights to aggregate the forecasts based on the past performance. Through a time-series cross-validation process, we empirically compare the performance of the newly developed ML-based frameworks with the legacy forecasting software and validate their potential value. The results demonstrate that the ML-based frameworks yield superior forecasting accuracy and improve on the two key performance measures related to forecast errors and biases. In addition, the required runtime decreases by approximately 70%, from more than 30 hours with the legacy software to approximately 8 hours using the ML-based system. The new forecasting system was fully deployed in early 2021 and has been regularly used to generate spare-parts demand forecasts worth more than 1 billion dollars. The ML pipelines have allowed Bombardier to improve its demand planning and inventory management operations, potentially translating into cost savings and increased customer satisfaction.

Acknowledgments

Bombardier thanks Traxxall for providing aerospace data related to aircraft maintenance in the exploratory phase. We also thank Karyne Turner and Thomas X Reid at Bombardier, as well as Soufiane El Mdari and Enrico Davide Cremonese at IVADO Labs for their contribution in project management and coordination. We gratefully acknowledge Florent Fettu and Vladimir Sonin at Bombardier who contributed to data preparation and the implementation of the ensemble model, as well as Morgan Bye, Guillaume Lepine, and Victor Lopez Selva at IVADO Labs who contributed to the development of ETL and machine learning pipeline during this project.

Appendix A. Flight Activity Features Construction

In this appendix, we describe how flight activity features are constructed. They capture the impact of flight activities on spare-parts maintenance over our entire modeling horizon.

A.1. Flight Activity Feature I: Hour-Landing Pattern

Hour-landing (HL) pattern is a categorical variable taking values from short, medium, or long ranges. It indicates the range of an airplane in which a spare part is installed. For example, a spare part with a long-range pattern implies this part is usually installed in aircraft that execute long-distance flight activities. To capture this information, we compute the feature, called the HL ratio, which can be calculated for spare part i at time t as follows:

$$HL\ Ratio_{i,t} = \frac{F_{i,t}}{L_{i,t}},$$

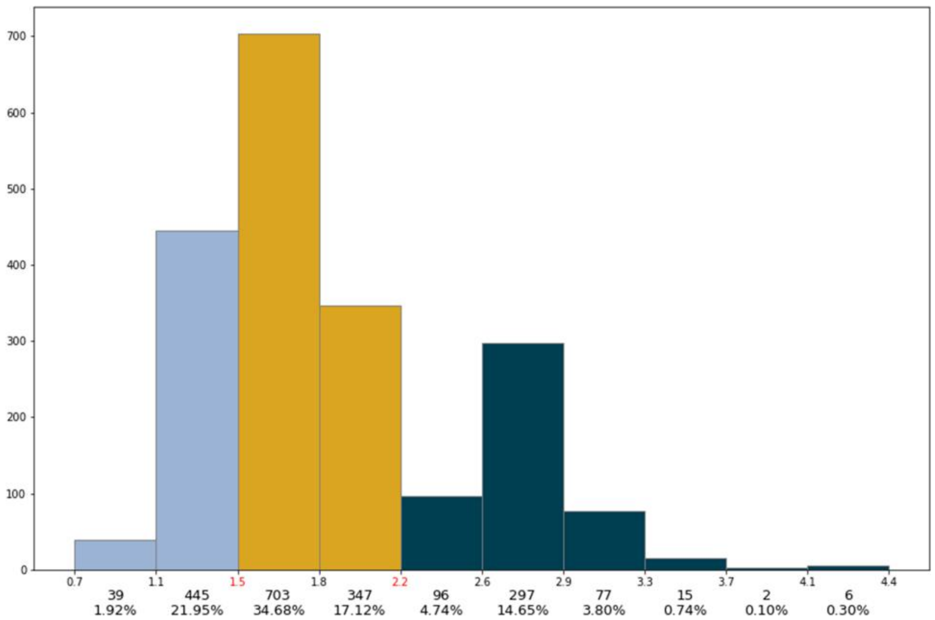
where $F_{i,t}$ is the cumulative flight hours of aircraft that have part i installed in month t, and $L_{i,t}$ is the cumulative number of landings of aircraft that have part i installed in month t.

Then, we calculate the average value of the HL ratios of each spare part over the time horizon of the data set (i.e., 2015 to 2019), which is used as a measure to categorize the parts. Finally, we cluster spare parts into short, medium, and long range based on the percentiles of their corresponding average HL ratios. More specifically, the parts that belong to the first 25th percentile, between the 25th and 75th percentiles, and the last 75th percentile are categorized as short, medium, and long range, respectively. We also verify the thresholds that correspond to the 25th and 75th percentiles using the data between 2010 and 2015 and find that these values and clusters remain static over time. Figure A.1 shows the distribution of spare-part HL ratios and the details of the three clusters.

A.2. Flight Activity Feature II: Flight Continuity

The *flight continuity* feature set represents the overall level of use of an aircraft part that depends on the inactivity and flight hours of each aircraft equipped with this specific part during a given month. More specifically, we define that a

Figure A.1. (Color online) Three HL Categories (Short, Medium, and Long) Marked by Different Shades that are Classified at the Thresholds of 1.5% and 2.2% in the Distribution Histogram



spare part i is considered *low activity* (mainly because it is nonflying) in month t if and only if the average monthly flight hours across aircraft that have part i installed ($f_{i,t}$) are less than a predefined threshold s. Based on this definition, we derive four features for each part i as follows:

- Proportion of months with low activity in which the monthly flight hours $f_{i,t}$ are lower than a predetermined threshold over the entire data horizon
- Average standard deviation of the monthly flight hours per month, which is the standard deviation of aggregated monthly flight hours $f_{i,t}$ divided by the total number of months in the modeling horizon
- Average low-activity month per aircraft, which equals the number of months classified as low activity divided by the number of aircraft equipped with this part in a given month
- Average standard deviation of aggregated monthly flight hours $f_{i,t}$ per aircraft divided by the number of aircraft equipped with this part in a given month

These features can be formally described as follows:

$$\text{Proportion of low-activity months of part } i = \frac{\sum_{t=1}^{N} I(f_{i,t} < s)}{N},$$

$$\text{Average standard deviation of flying times of part } i \text{ per month} = \frac{\sqrt{\sum_{t=1}^{N} (f_{t,i} - \bar{f}_i)^2}}{N},$$

$$\text{Average low-activity months per aircraft of of part } i = \frac{\sum_{t=1}^{N} I(f_{i,t} < s)}{U_i},$$

$$\text{Average standard deviation of flying times of part } i \text{ per aircraft} = \frac{\sqrt{\sum_{t=1}^{N} (f_{t,i} - \bar{f}_i)^2}}{U_i},$$

where $f_{i,t}$ is the flight hours of part i in month t, and s is a predetermined threshold. $I(f_{i,t} < s)$ is a binary vector where each value is equal to one if the condition is satisfied, $\bar{f}_i$ is the average number of flight hours over the whole modeling horizon, and U_i is the number of unique aircraft in Bombardier that have part i installed.

A.3. Flight Activity Feature III: Flight Behaviors

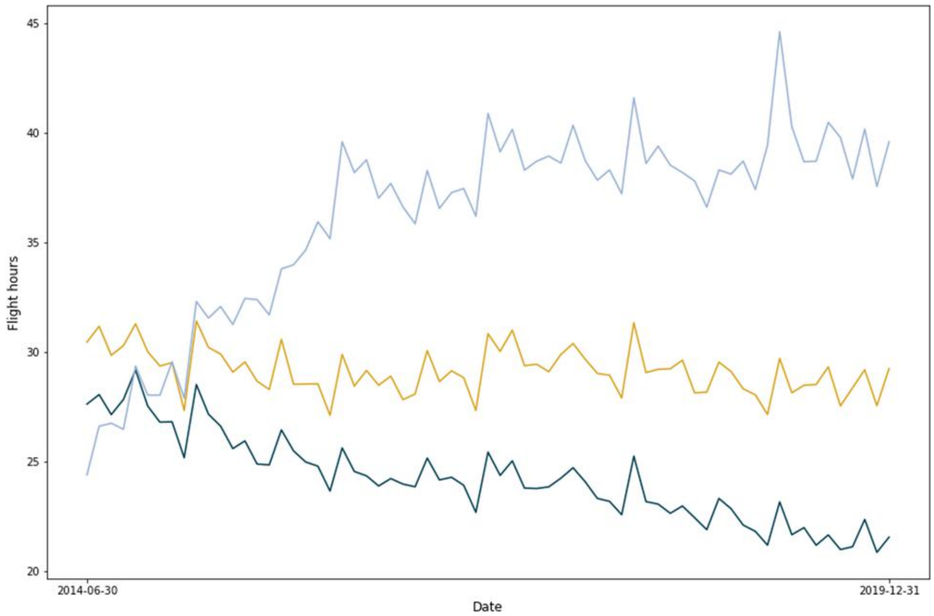
Besides flight range and continuity, we also cluster the flight behaviors of spare parts using time-series clustering methods. Note that standard clustering algorithms such as standard k-means are not suitable in this case because they do not take into account the sequence of time-series data points. In order to measure similarity among temporal sequences, we first derive the features that capture such sequences in time series using the soft dynamic time warping (soft-DTW) method (Cuturi and Blondel 2017). More specifically, we first aggregate the flight hours of each part i across regions and aircraft at the monthly level and apply the soft-DTW method. Once the transformed features based on soft-DTW are obtained, they can then be used in the standard clustering algorithms. In this step after feature transformation, we employ the k-means algorithm to cluster the parts into k clusters. Figure A.2 shows the mean flight hours of $k = 3$ clusters over the entire time horizon.

We would like to point out that the outputs from a survival analysis or a mean time to failure (MTTF) model, if available, can also be leveraged. This analysis is not in the scope of this study because of data limitation. The use of these potential features can be investigated in future work by the company using the analytics pipeline that has been developed.

Appendix B. Description of Input Features and Model Outputs

In the data pipeline, we prepare our input features at the part-region-month level before feeding to ML models. Table B.1

Figure A.2. (Color online) Mean Values of Three Different Clusters of Flight Hours of Aircraft



shows an illustrative example of the input features (X) and target (y) used in the machine-learning models (i.e., Elastic Net and LGBM), which correspond to the descriptions provided in Table B.1. Input features from multiple part numbers in the same group are concatenated together by rows. Some variables, such as region index (shown as "region" in Table B.1), are not time-varying, whereas the variables that correspond to the part age and flight activity are aggregated until the month that corresponds to a specific data point. To incorporate these categorical features in the LGBM model, we can simply indicate that these features are categorical in the LGBM library, and it automatically transforms the features using the integer-encoded categorical features (Microsoft Corporation 2022). In the Elastic Net model, these categorical variables are encoded as dummy variables prior to being used in the model.

Similarly, model outputs follow the same index structure and produce monthly forecasts for each part number in each region. Table B.2 shows an example of the structure of model outputs. In the live monthly forecast in the production pipeline, we feed input features which are required for the forecasts to the trained models. Most forecasts are produced for the next month $l = 1$. In an exceptional case for some parts with the supply lead time longer than one month, the team can also produce the next l month forecast by training the machine-learning models for a specific look-ahead period l separately using the same pipeline. In addition, the resulting forecasting errors (i.e., mean absolute deviation) determined in the cross-validation process are used to determine the level of safety stocks in the inventory fulfillment system (Krupp 1997).

We further note that, for the traditional time series models (e.g., Croston's methods), the models simply use the historical demands of the period prior to the period to forecast (i.e., data in column "Demand" in Table B.1 as in traditional time series models). Thus, the input features columns are not required for these models.

Appendix C. ML-Based Algorithm for Demand Forecasting and Methodologies for Hyperparameter Tuning

C.1. Feature Engineering

The two parameters used in the features engineering process are m (the length of demand lags) and n (the length of prior periods used to calculate demand statistics). In our pipeline, we set the sets of both m and n values to $\{2, 4, 6, 12\}$ (these sets can be modified by the demand planner if necessary). We determine the best values of m and n using a validation set. In our experiments, the best values of m and n are 4 and 6 for rotable parts and 6 and 12 for expendable parts, respectively.

C.2. scikit-Learn Library for Elastic Net Regression

The `scikit-learn` library is used to train the Elastic Net model using the model object `sklearn.linear_model.ElasticNet()`. We use the default parameters in our implementation with one important parameter, `l1_ratio`, which assigns the weight to L1 (lasso) and L2 (ridge) regularizations. In our hyperparameter search procedure, a parameter-tuning process can be performed using a step-size approach to determine an optimal parameter in the range $0 \leq$ `l1_ratio` ≤ 1.

C.3. LGBM Library

We implement the LGBM model by using the LGBM library developed by Microsoft Corporation (2022). This library is widely used by researchers and practitioners (Makridakis et al. 2022). The overall design and use of the LGBM is aligned with the widely adopted `scikit-learn` machine-learning library (Pedregosa et al. 2011). In this LGBM library, feature importance

Table B.1. Illustrative Example of Input Features for ML-Based Models (LGBM and Elastic Net)

Demand data				Input features (X)															Target (y)
				Demand lags/statistics/intermittency									Region/month index		Flight and other features				
Part no.	Region	Time	Demand	lag-1	…	lag-m	Rolling Mean	…	ADI	…	Season (cos)	…	Region[a]	Month[a]	HL pattern[a]	…	Part age		$t+l$ ($l=1$)
111	AMERICA	Sep-2018	9	7		0	7.8		1.125		0.000		AMERICA	Sep	High		7.679		9
	AMERICA	Oct-2018	5	9		4	8.1		1.143		0.500		AMERICA	Oct	High		7.679		5
	AMERICA	Nov-2018	3	5		5	6.7		1.024		0.866		AMERICA	Nov	High		8.042		3
	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮		⋮		⋮	⋮	⋮	⋮	⋮		⋮
	EMEA	Sep-2018	3	0		0	0.9		1.813		0.000		EMEA	Sep	High		7.679		3
	EMEA	Oct-2018	1	3		1	1.8		1.734		0.500		EMEA	Oct	High	⋮	7.679		1
	EMEA	Nov-2018	2	1		2	1.6		1.456		0.866		EMEA	Nov	High		8.042		2
	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮		⋮		⋮	⋮	⋮	⋮	⋮		⋮
	APAC	Sep-2018	4	0		5	1.2		2.825		0.000		APAC	Sep	High		7.679		4
	APAC	Oct-2018	0	4		2	1.8		2.250		0.500		APAC	Oct	High		7.679		0
	APAC	Nov-2018	2	0		0	1.7		2.817		0.866		APAC	Nov	High		8.042		2
	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮		⋮		⋮	⋮	⋮	⋮	⋮		⋮
124	AMERICA	Sep-2018	0	1		2	1.6				0.000		AMERICA	Sep	Med				0
	⋮	⋮	4	0	⋮	1	⋮				⋮		⋮	⋮	⋮				⋮

Notes. This table follows the list of features provided in Table 1. Some columns and data that belong to the same feature type or column are omitted (as indicated by …) due to limited space.
[a]Categorical features.

and plot prediction trees could be directly obtained through the `feature_importance()` function. In this project, we opt for the use of default parameters to simplify the overall retraining process, which will be carried on by the demand planning team. For the full list of hyperparameters of the LGBM library, we refer the interested reader to LGBM documentation (Microsoft Corporation 2022), which also discusses important parameter configurations under the section *Parameters Tuning*. To allow such flexibility, we also allow the planner to optimize some important hyperparameters, including learning rate, which can be between zero and one (the value is 0.1 by default)—and the loss function, which can be changed to either L1 loss or L2 loss (currently used by default), and Tweedie regression.

Appendix D. Forecast Performance Measures

This section describes the two performance measures used in this study: root-mean-squared scaled error (RMSSE) and normalized forecasting bias (NFB).

D.1. RMSSE

Relative forecast error measures of time series, such as mean absolute percentage error, can potentially result in an extremely high value (or even infinity) when measuring slow-moving demands because of division by zero or a value close to zero. Thus, the common measures based on relative errors are unsuitable for measuring intermittent demand forecast, wherein many consecutive zero demands tend to occur in different time periods. To resolve this issue, Hyndman and Koehler (2006) propose the MASE measure and RMSSE as a variant of the MASE, in which the difference between forecasts and actual values are measured using root-mean-square errors rather than absolute deviations as in MASE. Specifically, RMSSE calculates the out-of-sample averaged squared error in the numerator, which is divided by an in-sample one-step-forward averaged error. This RMSSE measure can be formally stated as follows:

$$RMSSE = \sqrt{\frac{1}{h}\frac{\sum_{t=N+1}^{N+h}(y_t - \hat{y}_t)^2}{\frac{1}{N-1}\sum_{t=2}^{N}(y_t - y_{t-1})^2}},$$

where y_t is the actual demand at time t, $\hat{y}_t$ is the forecast at time t, N is the number of in-sample observations (number of historical observations), and h is the forecasting horizon. In our case, we compute the denominator of RMSSE only for the time periods when spare parts are actively sold—that is, the periods following the first observed nonzero demand for the time series under evaluation.

D.2. NFB

We adapt this bias measurement from the tracking signal (TS) proposed by Alstrøm and Madsen (1996). TS measures the performance and bias of forecasts based on cumulative errors. The cumulative errors are related to inventory management performance (e.g., a large negative cumulative error can lead to overstock of parts inventory). TS is commonly defined as the ratio of the cumulative sum of forecasting errors (i.e., the difference between the forecasts and the actual values) to the standard deviation (or mean absolute deviation) of errors (Alstrøm and Madsen 1996).

Table B.2. Example of Outputs of Forecasting Models

Part/region/month			Forecast ($\hat{y}$)
Part no.	Region	Date	$t+1$
111	AMERICA	Jan-2019	7.488
		Feb-2019	3.367
		Mar-2019	1.245
		...	...
	EMEA	Jan-2019	2.385
		Feb-2019	0.324
		Mar-2019	1.946
		...	...
	APAC	Jan-2019	1.635
		Feb-2019	0.828
		Mar-2019	1.423
		...	...
124	AMERICA	Jan-2019	0.875
	....	...	...

Note. Some data in the same column are omitted (indicated by ...).

To use this type of measure to compare the performances of forecasts from different models, rather than using the standard deviation of forecasting errors of each corresponding model, we set the denominators of all the models to the standard deviation of errors of forecasts produced by a reference model σ_e^r (which are the forecasts generated by the legacy system in our case). For each part-region level, NFB is calculated as follows:

$$NFB = \frac{\sum_{t=1}^{h} e_t}{\sigma_e^r},$$

where $e_t = y_t - \hat{y}_t$, $\sigma_e^r = \sqrt{\frac{1}{h}\sum_{t=1}^{h}(e_t^r - \bar{e}^r)^2}$, and y_t is the actual demand at time t, $\hat{y}_t$ is the forecast demand at time t, e_t is the forecasting error at time t, and e_t^r is the forecasting error at time t of the forecasts created by the reference (benchmark) model; h is the length of the reference forecasting period to determine the standard deviation of the forecast errors and $\bar{e}^r$ is the average value of e_t^r during the reference period of length h. The numerator of NFB captures the cumulative forecasting errors over horizon h for each part-region combination. We scale (normalize) this value by the standard deviation of the forecasting errors of the reference model.

Appendix E. Feature Importance Obtained from LGBM

Apart from forecasting accuracy, exploring which input features contribute the most to the forecast also brings insights into our spare-parts demand forecasting. Because LGBM is a tree-based machine-learning algorithm, feature importance can be directly obtained from the model. This feature importance represents the relative importance score of each of the input features that the algorithm used to produce the forecast.

Figure E.1 shows the 30 most important features that contribute to the forecasts of the LGBM model based on RML framework (which include flight activity data). The value on the x axis serves only a comparative purpose among different input features (the higher the value, the more important the input feature is). There are in total 35 input features used by the LGBM model in RML framework, 30 of which are output in this figure. The top three features are *ADI* and CV^2, which capture intermittent characteristics, as well as the current month demand prediction from our model. The two supplementary features, part age and the number of unique aircraft, are the fourth and the sixth most important feature,

Figure E.1. (Color online) Thirty Most Important Input Features Based on the LGBM Model for Rotable Parts

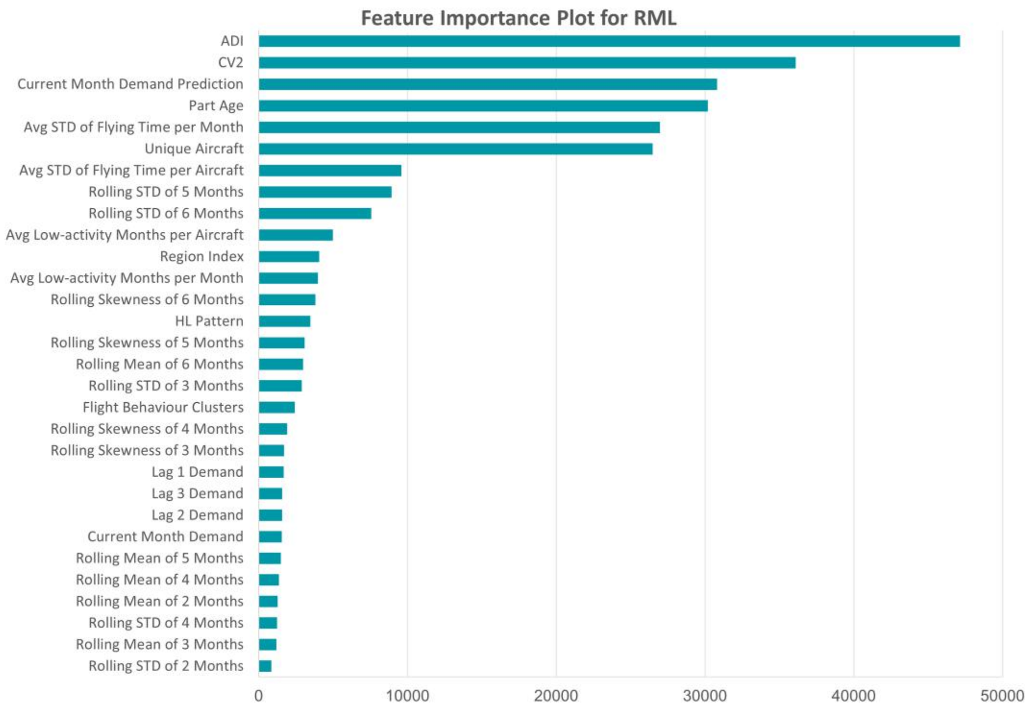
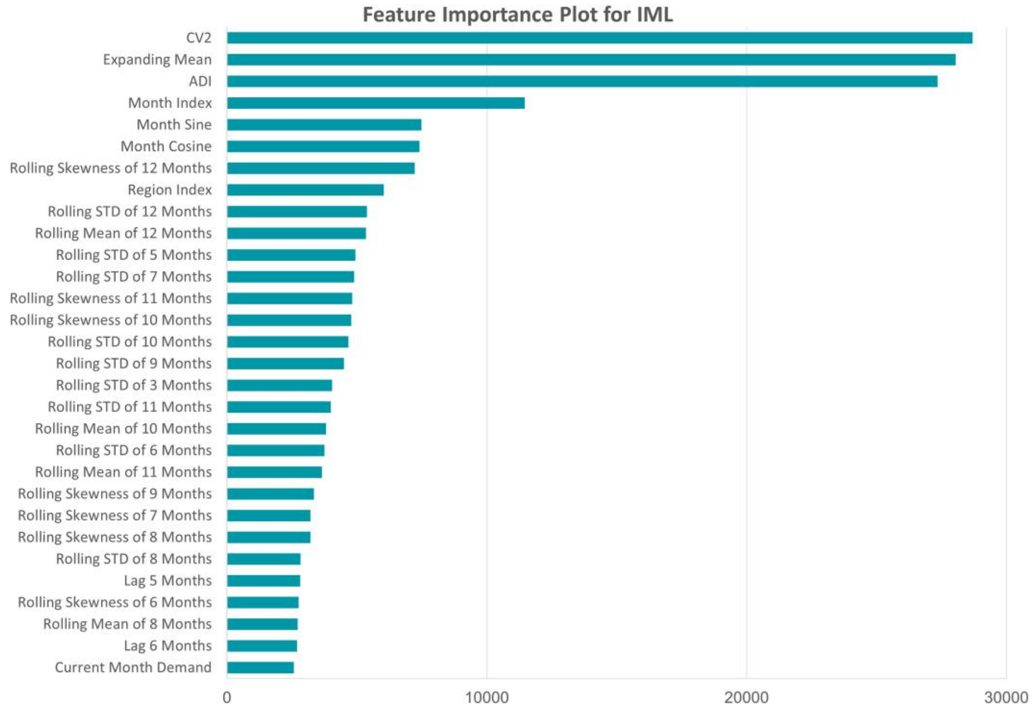


Figure E.2. (Color online) Thirty Most Important Input Features Based on the LGBM Model for Expendable Parts



respectively. Meanwhile, the flight activity features that capture flight continuity rank at 5th, 7th, 10th, and 12th. Other flight activity features, including HL pattern and flight behaviors, are identified as the top important features. The high ranking of our flight activity features proves that the feature engineering process has successfully associated flight activity with the worn-out speed of spare parts. Figure E.2 presents the 30 most important features out of 57 total input features for the LGBM model of IML framework (the flight activity data are not sufficiently available for this group, as described earlier in the main paper). Similarly, intermittent characteristic features rank at the top of our model. The seasonality components (month sine and month cosine) also show an important contribution to forecasting accuracy. Among the demand statistics features, lag demand, and rolling demand features over a six-month period or longer are considered more important features by the LGBM model for expendable parts.

References

Allison PD (1999) *Multiple Regression: A Primer* (Pine Forge Press, Thousand Oaks, CA).

Alstrøm P, Madsen P (1996) Tracking signals in inventory control systems a simulation study. *Internat. J. Production Econom.* 45(1–3):293–302.

Amin-Naseri MR, Tabar BR (2008) Neural network approach to lumpy demand forecasting for spare parts in process industries. *Proc. Internat. Conf. on Comput. and Comm. Engrg.* (IEEE, Piscataway, NJ), 1378–1382.

Amirkolaii KN, Baboli A, Shahzad M, Tonadre R (2017) Demand forecasting for irregular demands in business aircraft spare parts supply chains by using artificial intelligence (ai). *IFAC-PapersOnLine* 50(1):15221–15226.

Bajari P, Nekipelov D, Ryan SP, Yang M (2015) Demand estimation with machine learning and model combination. Technical report, National Bureau of Economic Research, Cambridge, MA.

Bates JM, Granger CW (1969) The combination of forecasts. *J. Oper. Res. Soc.* 20(4):451–468.

Bergmeir C, Hyndman RJ, Koo B (2018) A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Statist. Data Anal.* 120:70–83.

Boylan JE, Syntetos AA, Karakostas GC (2008) Classification for forecasting and stock control: A case study. *J. Oper. Res. Soc.* 59(4):473–481.

Cerqueira V, Torgo L, Pinto F, Soares C (2019) Arbitrage of forecasting experts. *Machine Learn.* 108(6):913–944.

Clemen RT (1989) Combining forecasts: A review and annotated bibliography. *Internat. J. Forecasting* 5(4):559–583.

Croston JD (1972) Forecasting and stock control for intermittent demands. *J. Oper. Res. Soc.* 23(3):289–303.

Cuturi M, Blondel M (2017) Soft-dtw: A differentiable loss function for time-series. Precup D, Teh YW, eds. *Proc. Internat. Conf. on Machine Learn.* (ICML, San Diego, CA), 894–903.

Ferreira KJ, Lee BHA, Simchi-Levi D (2016) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing Service Oper. Management* 18(1):69–88.

Gaillard P, Goude Y (2020) opera: Online prediction by expert aggregation. Accessed October 1, 2022, https://CRAN.R-project.org/package=opera.

Gamberini R, Lolli F, Rimini B, Sgarbossa F (2010) Forecasting of sporadic demand patterns with seasonality and trend components: An empirical comparison between Holt-Winters and (S)ARIMA methods. *Math. Problems Engrg.* 2010:1–14.

Goehry B, Goude Y, Massart P, Poggi JM (2019) Aggregation of multi-scale experts for bottom-up load forecasting. *IEEE Trans. Smart Grid* 11(3):1895–1904.

Gutierrez RS, Solis AO, Mukhopadhyay S (2008) Lumpy demand forecasting using neural networks. *Internat. J. Production Econom.* 111(2):409–420.

Hewamalage H, Bergmeir C, Bandara K (2021) Recurrent neural networks for time series forecasting: Current status and future directions. *Internat. J. Forecasting* 37(1):388–427.

Hyndman RJ, Athanasopoulos G (2018) *Forecasting: Principles and Practice* (OTexts).

Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. *Internat. J. Forecasting* 22(4):679–688.

Januschowski T, Wang Y, Torkkola K, Erkkilä T, Hasson H, Gasthaus J (2021) Forecasting with trees. *Internat. J. Forecasting* 38(4):1473–1481.

Johnston F, Boylan JE (1996) Forecasting for items with intermittent demand. *J. Oper. Res. Soc.* 47(1):113–121.

Johnston F, Boylan JE, Shale EA (2003) An examination of the size of orders from customers, their characterisation and the implications for inventory control of slow moving items. *J. Oper. Res. Soc.* 54(8):833–837.

Jose VRR, Winkler RL (2008) Simple robust averages of forecasts: Some empirical results. *Internat. J. Forecasting* 24(1):163–169.

Karthikeswaren R, Kayathwal K, Dhama G, Arora A (2021) A survey on classical and deep learning based intermittent time series forecasting methods. *Proc. Internat. Joint Conf. on Neural Networks* (IEEE, Piscataway, NJ), 1–7.

Kaya GO, Sahin M, Demirel OF (2020) Intermittent demand forecasting: A guideline for method selection. *Sadhana* 45(1):1–7.

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inform. Processing Systems*, vol. 30 (NeurIPS, San Diego).

Kolassa S (2011) Combining exponential smoothing forecasts using akaike weights. *Internat. J. Forecasting* 27(2):238–251.

Kourentzes N (2013) Intermittent demand forecasts with neural networks. *Internat. J. Production Econom.* 143(1):198–206.

Krupp JA (1997) Safety stock management. *Production Inventory Management J.* 38(3):11.

Levén E, Segerstedt A (2004) Inventory control with a modified croston procedure and erlang distribution. *Internat. J. Production Econom.* 90(3):361–367.

Makridakis S, Winkler RL (1983) Averages of forecasts: Some empirical results. *Management Sci.* 29(9):987–996.

Makridakis S, Spiliotis E, Assimakopoulos V (2022) M5 accuracy competition: Results, findings, and conclusions. *Internat. J. Forecasting* 38(4):1346–1364.

Microsoft Corporation (2022) LightGBM's documentation and Python API. Accessed May 20, 2022, https://lightgbm.readthedocs.io.

Niu Y (2020) Walmart sales forecasting using xgboost algorithm and feature engineering. *Proc. Internat. Conf. on Big Data & Artificial Intelligence & Software Engrg.* (IEEE, Piscataway, NJ), 458–461.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, et al. (2011) Scikit-learn: Machine learning in Python. *J. Machine Learn. Res.* 12:2825–2830.

Petropoulos F, Kourentzes N (2015) Forecast combinations for intermittent demand. *J. Oper. Res. Soc.* 66(6):914–924.

Pollock D (1993) *Lectures in Time-Series Analysis and Forecasting* (Queen Mary & Westfield College, London).

Porras E, Dekker R (2008) An inventory control system for spare parts at a refinery: An empirical comparison of different re-order point methods. *Eur. J. Oper. Res.* 184(1):101–132.

Rao AV (1973) A comment on: Forecasting and stock control for intermittent demands. *J. Oper. Res. Soc.* 24(4):639–640.

Schultz CR (1987) Forecasting and inventory control for sporadic demand under periodic review. *J. Oper. Res. Soc.* 38(5):453–458.

Snyder R (2002) Forecasting sales of slow and fast moving inventories. *Eur. J. Oper. Res.* 140(3):684–699.

Snyder RD, Ord JK, Beaumont A (2012) Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *Internat. J. Forecasting* 28(2):485–496.

Strijbosch L, Heuts R, Van der Schoot E (2000) A combined forecast-inventory control procedure for spare parts. *J. Oper. Res. Soc.* 51(10):1184–1192.

Syntetos AA, Boylan JE (2001) On the bias of intermittent demand estimates. *Internat. J. Production Econom.* 71(1–3):457–466.

Syntetos AA, Boylan JE (2005) The accuracy of intermittent demand estimates. *Internat. J. Forecasting* 21(2):303–314.

Syntetos AA, Boylan JE (2006) On the stock control performance of intermittent demand estimators. *Internat. J. Production Econom.* 103(1):36–47.

Syntetos AA, Boylan JE, Croston J (2005) On the categorization of demand patterns. *J. Oper. Res. Soc.* 56(5):495–503.

Teunter RH, Syntetos AA, Babai MZ (2010) Determining order-up-to levels under periodic review for compound binomial (intermittent) demand. *Eur. J. Oper. Res.* 203(3):619–624.

Teunter RH, Syntetos AA, Babai MZ (2011) Intermittent demand: Linking forecasting to inventory obsolescence. *Eur. J. Oper. Res.* 214(3):606–615.

Willemain TR, Smart CN, Schwarz HF (2004) A new approach to forecasting intermittent demand for service parts inventories. *Internat. J. Forecasting* 20(3):375–387.

Willemain TR, Smart CN, Shockor JH, DeSautels PA (1994) Forecasting intermittent demand in manufacturing: A comparative evaluation of Croston's method. *Internat. J. Forecasting* 10(4):529–538.

Williams T (1984) Stock control with sporadic and slow-moving demand. *J. Oper. Res. Soc.* 35(10):939–948.

Zhang Y, Tiňo P, Leonardis A, Tang K (2021) A survey on neural network interpretability. *IEEE Trans. Emerging Topics Comput. Intelligence* 5(5):726–742.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J. Royal Statist. Soc. Ser. B Statist. Methodology* 67(2):301–320.

Verification Letter

Patrick Lemieux, Director, Parts Operations & Inventory Planning, Bombardier, 400 Côte-Vertu Road West, Dorval, Quebec H4S 1Y9, Canada, writes:

"It is my pleasure to provide a verification letter for the research article 'Bombardier Aftermarket Demand Forecast in Business Aircraft Spare Parts with Machine Learning.' This article has been written based on the joint initiative between Bombardier Aerospace and our industrial partner IVADO Labs. The main objective of this collaboration is to improve the demand forecasting process and performance of aftermarket spare parts at Bombardier which was originally performed using a legacy forecasting system.

"This project started in September 2020 and was completed in February 2021. The team has designed, implemented, and deployed an AI-based forecasting system that yield significant improvements in terms of forecasting performance and runtime scalability. In the validation phase of the project prior to deployment, the new AI-based system has improved the overall accuracy and normalized biases, which are our key performance measures, by approximately 7% and 5%, respectively. The team has then successfully deployed this new system in February 2021, and it is being used regularly to create monthly forecasts for our aftermarket demand and inventory planning since then. Throughout February to July 2021 after deployment, Bombardier Aftermarket successfully improved the overall forecasting accuracy by approximately 5.3% with an overall unbiased forecast of 75.8% by using this new AI-based system. This new process also saves more than 70% of forecasting runtime compared with the legacy software, which requires approximately 30 hours to complete the process. These results translate into more efficient and responsive inventory planning as well as improved customer satisfaction.

"Please do not hesitate to contact me if you require further information."

Pierre Dodin holds a PhD in mathematics and operational research from University Pierre et Marie Curie. He has applied mathematical modeling and machine learning to applications in target tracking and object classification. Previously at ArthroLab, he helped design the technology to segment MRI images, which became part of several medical papers proving osteoarthritis treatment efficiency. He serves as a principal data scientist at Bombardier with a focus on spare-parts demand prediction.

Jingyi Xiao holds an MS degree in data science and business analytics from HEC Montréal. Her interests lie in tree-based machine-learning algorithms, time series forecasting, and cloud data engineering. Additionally, she holds a financial risk manager certification, which enhances her ability to implement quantitative models in risk management. Currently, she is a senior consultant at Omnia AI, Deloitte Canada, where her focus is on cloud services and data engineering.

Yossiri Adulyasak is the Canada research chairholder in supply chain analytics and an associate professor at HEC Montréal. Prior to that, he was a postdoc at Massachusetts Institute of Technology and a data scientist at a global software company. He coauthored more than 35 publications in supply chain AI/analytics and optimization that appeared in prominent research journals. He serves as a scientific consultant for multiple firms and is a coinventor of five patent applications.

Neda Etebari Alamdari holds a PhD in applied mathematics from Polytechnique Montréal. She is a highly accomplished data scientist with more than eight years of experience in various artificial intelligence (AI)-related industrial and academic roles. Her publications in AI and optimization have been featured in a number of prominent journals and conferences.

Lea Gauthier is a data scientist lead at IVADO Labs with more than 10 years of experience in artificial intelligence (AI) development across multiple sectors. Previously, she worked on algorithm development in finance, advertising, and gaming sectors. She holds a PhD from the University Paris-Sud in experimental particle physics and was involved in the ATLAS experiment at the CERN. After graduation, she was a postdoctoral researcher at the University of Montreal.

Philippe Grangier completed a PhD in operations research at IMT Atlantique in 2015. Part of his thesis work was awarded the 2017 Practice Prize Award by the Canadian Operational Research Society. He was later a Mitacs Elevate postdoctoral scholar at Polytechnique Montréal and JDA Software (now Blue Yonder). Following this, he has worked as data scientist on applied optimization and machine learning projects. He is now a senior data scientist at IVADO Labs.

Paul Lemaitre is currently a data scientist at IVADO Labs, a Montreal-based artificial intelligence consulting firm. After an industrial PhD with a focus on statistics, probabilities, and machine learning, he worked in various industries and research institutions such as McGill University and Électricité de France research and development on domains encompassing heavy chemistry, nuclear safety, and neuroimaging.

William L. Hamilton is an adjunct professor of computer science at McGill University and a senior quantitative researcher at Citadel LLC. His main research is on the development of machine learning models that can reason about the complex and interconnected world with particular interests in graph representation learning and natural language processing. He is the author of the book Graph Representation Learning and has published more than 40 research articles in machine learning and AI.