Closed Loop Supply Chain Waste Reduction Through Predictive Modelling and Process Analysis

by

Hans P. Kobor

B.S. Mechanical Engineering, United States Military Academy, 2011

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

Master of Business Administration
and
Master of Science in Mechanical Engineering

In conjunction with the Leaders for Global Operations Program at the
Massachusetts Institute of Technology
June 2019

Signature of Author _____ Signature redacted
Mechanical Engineering
MIT Sloan School of Management
May 7, 2019

Certified by _____ Signature redacted
Juan Pablo Vielma, Thesis Supervisor
Richard S. Leghorn (1939) Career Development Professor, MIT Sloan School of Management

Certified by _____ Signature redacted
Duane Boning, Thesis Supervisor
Clarence J. LeBel Professor, Electrical Engineering and Computer Science

Certified by _____ Signature redacted
Sanjay Sarma, Thesis Reader
Fred Fort Flowers (1941) and Daniel Fort Flowers (1941) Professor of Mechanical Engineering

Accepted by _____ Signature redacted
Nicolas Hadjiconstantinou
Chair, Mechanical Engineering Committee on Graduate Students
Department of Mechanical Engineering

Accepted by _____ Signature redacted
Maura Herson
Assistant Dean, MBA Program, MIT Sloan School of Management

This page intentionally left blank.

Closed Loop Supply Chain Waste Reduction Through Predictive Modelling and Process Analysis

by

Hans Kobor

Submitted to the MIT Sloan School of Management and the Department of Mechanical Engineering on May 7, 2019 in Partial Fulfillment of the Requirements for the Degrees of Master of Business Administration and Master of Science in Mechanical Engineering.

Abstract

Verizon distributes Customer Premises Equipment (CPE) such as set top boxes, broadband routers, and WiFi extenders to Fios customers via a variety of paths; for example: direct ship to customer (either for self-install or for later installation by a field technician), delivery via field technicians, or retail store pickup (primarily for self-install). Each method has its own benefits and shortcomings due to impacts on metrics such as inventory levels, shipping costs, on-time delivery, and system complexity. Although the majority of shipments are successfully activated in the customer's home, a non-trivial percentage results in unused returns or inventory shrinkage. These undesirable results represent a significant amount of wasted resources. This thesis is focused on identifying and realizing cost savings in the Fios supply chain through reduction in waste associated with unsuccessful shipments.

In order to effectively analyze the closed-loop supply chain, accurate and reliable process mapping is critical. Interviews with key stakeholders, together with order and shipment data analysis yielded a complete picture of the ecosystem's processes and infrastructure. Process mining techniques augmented this understanding, using event log data to identify and map equipment and information flows across the supply chain. All together this analysis is used to identify order cancellations as a key source of waste.

To limit waste, it is necessary to conduct analysis both internal to Verizon's processes and externally, to determine if there are customer trends leading to order termination. Process mining was used for the internal analysis and, while it helped identify singular cases in which process abnormalities were associated with undesirable outcomes, its current form proved unsuited for root cause analysis. Internal analysis did, however, illuminate opportunities for improvement in radio-frequency identification (RFID) usage and protocols across the supply chain. Current systems can result in poor visibility of equipment as it moves within some segments of the supply chain. The actual monetary impact is difficult to determine but likely to increase as the importance of RFID increases.

External analysis is conducted through predictive modelling. Using a variety of data sources, a model with over 80% sensitivity and a low false positive rate is achieved. Operationalizing this model through real time incorporation with sales was explored but found to be overly complex. Instead, the random forest model yielded policy changes guided by the features with the highest importance. A pilot is currently in development to test the efficacy of suggested changes, as the model implies significant savings opportunity.

Thesis Supervisor: Duane Boning
Clarence J. LeBel Professor, Electrical Engineering and Computer Science

Thesis Supervisor: Juan Pablo Vielma
Richard S. Leghorn (1939) Career Development Professor, MIT Sloan School of Management

This page intentionally left blank.

Acknowledgments

First and foremost, I would like to express my gratitude to the entire Verizon Global Supply Chain organization for their support and collaboration throughout the project, especially my supervisor, Steve Baum. He facilitated this project as a rich learning experience; his shared expertise and candid discussion helped me grow both personally and professionally. Sincerest thanks also go to my colleagues in the group, particularly Ernesto Allwood, and Jen Canlas. They were instrumental in helping me retrieve, contextualize, and process mountains of data from myriad sources. Special thanks also go to Alan Rompala, Sam Mastruserio, Matt Moore, and Frank Frontiera. Their guidance and eagerness to share knowledge provided me with the environment and tools to succeed.

I would also like to thank my faculty advisors, Duane Boning and Juan Pablo Vielma. Each was a wonderful source of knowledge and ideas, and their guidance was instrumental in shaping the direction of the project.

The LGO staff, particularly Patty Eames and Ted Equi, also deserve acknowledgement for their support and guidance during not only the internship, but also the entire LGO program.

Finally, I would like to thank my wife Lauren for her tremendous love and support. Her monumental strength and ability to manage the kids while we were apart allowed me to stay focused and succeed on the project and throughout the program. She is my rock, and I could not have done this without her.

This page intentionally left blank.

Table of Contents

List of Figures

List of Tables

1 Introduction

This thesis presents techniques to analyze and improve the efficiency of a closed loop supply chain and reduce the waste associated with certain operational decisions and technologies. The thesis presents a predictive model that can be employed to limit unused and wasteful shipments. Model discussion depicts a solution to optimize parameters for a low false positive rate despite a highly class-imbalanced data set. In addition, this thesis presents a technique to understand complex system interactions and process flows. Finally, we discuss device tracking technology within the supply chain and its operational usage and impact. This chapter provides a brief overview of Verizon as a whole, and of Fios and its supply chain operations. Additionally, this chapter provides an overview of potentially problematic features within the system, and the methodologies with which the project was approached.

1.1 Verizon Background

Verizon is an American multinational conglomerate and a global leader in communications. The company was founded in 2000 as a merger between Bell Atlantic and GTE Corporation. Over the last 19 years the company has grown through dozens of mergers and acquisitions with some of the most prominent names in American technology, including Yahoo, MCI, AOL, and Alltel, among others. Today, Verizon is a leading provider of communications, information, and entertainment products and services to consumers, businesses, and government agencies worldwide. In 2018, the company earned $11.6 Billion in EBITDA (earnings before interest, taxes, and depreciation) on $130.9 billion in total revenue. At the time of this project, the company was divided into three main segments: Wireless, Wireline, and Enterprise.

- Wireless – provides wireless voice and data services, as well as equipment sales
- Wireline – broadband video, voice, and data, corporate networking solutions, data center, and cloud services for residences and small businesses.
- Enterprise – provides cloud-based solutions for corporate and government entities, delivering security, mobility, and information-sharing solutions.

This organizational structure has since been changed, effective January 1, 2019 to enable the business to focus more directly on the rollout and deployment of 5G networks across the United States. The new structure divides the business into three new segments: Consumer, Business, and Media.

- Consumer – includes the consumer segment of the company's wireless and wireline businesses
- Business – includes products and services sold to businesses and government from both the wireless and wireline businesses

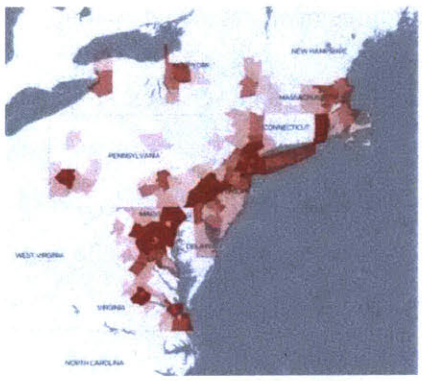
- Media – based on the Oath Group and its media properties, this segment focuses on generating and disseminating media content, advertising, and technology

Although it is important to recognize the changing dynamics of the business, for the purpose of this document, analysis is conducted according to the old wireless/wireline structure to avoid confusion.

1.2 Verizon Fios

Verizon's Fios service provides television, internet, and voice services to residential and small business customers in the mid-Atlantic and northeastern United States through a 100% fiber optic network, as illustrated in Figure 1-1 [2]. Fios serves about 6 million internet and 4.5 million TV subscribers and is a major portion of the wireline business, accounting for about 40% of the unit's total revenue. Fios marketing emphasizes high data transfer speeds, excellent reliability, and best-in-segment customer service. Despite these excellent qualities, Fios overall subscription growth has slowed in recent years to about 2.5% per year, and that growth is entirely from internet services. TV subscription has experienced a net decrease in customers over the past two years, as streaming services such as Netflix and Hulu have taken their toll on the Cable TV industry [1]. In addition to this slow growth rate, in 2013, Verizon sold its Fios network assets in California, Texas, and Florida. As a result, Fios operates in a relatively limited service area.

Figure 1-1: Fios Service Area



Given the decision to limit the geographical service area of Fios and the lack of substantial organic growth, the business unit is considered to be in the mature portion of its lifecycle [2]. As a result, controlling costs is a key metric within the business, and this thesis focuses on costs within the Fios supply chain in particular.

1.3 Wireline/Fios Supply Chain

Fios services require certain pieces of equipment at the customer's home to ensure network functionality and a positive customer experience. For a typical "triple-play" bundle customer (the most common type), there are three main pieces of required equipment: Optical Network Terminal (ONT), Router, and Set-top box. The ONT is the interface between the fiber optic network and the in-home network. The router receives input from the ONT via coaxial or Ethernet cable and transfers information between the ONT and in-home devices via either wi-fi or Ethernet. The set-top box receives television signal from either the ONT or router (depending on device) and provides basic TV viewing, with the potential for adding DVR capabilities.

An ONT is required in any house to enable Fios services, and customers are required to rent Verizon-provided set-top boxes as part of their TV subscription. Customers are able to provide their own router and abstain from renting or purchasing a Verizon router; however, any customer with internet speeds higher than 100 Megabytes per second is required to rent or purchase the newest router, the BHR4. Fios customers are also encouraged to purchase or rent wi-fi network extenders (FNEs) to ensure quality internet coverage throughout the home. Although there are only four device types (ONT, STB, Router, FNE), there are multiple legacy devices that are prevalent across the network, resulting in over 20 managed SKUs. The wireline supply chain group manages the forward and reverse logistics for all of these devices and plays a critical role in ensuring a quality customer experience at installation and in the event of needed timely replacements or repairs.

In order to ensure proper distribution and delivery, the supply chain operates three distinct channels for forward logistics:

- Retail: Customers may pick up or purchase devices at Verizon retail stores
- Garage Work Centers (GWCs): Installation and repair technicians work out of GWCs, which maintain a small inventory of most SKUs. Technicians frequently pull devices out of GWC inventory to complete both installation and repair orders.
- Direct to consumer (DTC): Many customers receive direct shipments of equipment from the regional distribution center (RDC) to their home. This is a frequent channel for both repair and installation orders. For installation orders, the devices are shipped to the customer and then a technician arrives on the scheduled date to conduct the installation. This arrangement is known as "Direct Ship." DTC repair orders are typically overnighted

to the customer with a return box for the defective device. This is known as a "dropship" order.

Each of these distribution channels is operational for reverse logistics as well. Customers frequently drop equipment off at retail stores or mail devices back to the RDC. Technicians also regularly bring defective equipment to the GWC for return to the RDC.

The focal point for the Fios supply chain is the RDC. There are two such facilities and each are located in eastern Pennsylvania. They are operated by separate third party logistics providers (3PLs). Each RDC receives bulk shipments from manufacturers and suppliers, which it allocates through the aforementioned distribution channels. The RDCs process over 500,000 shipments and returns per month. They also maintain buffer stock inventory for the supply chain. The eastern Pennsylvania location is centralized within the Fios service area and allows for overnight shipments of equipment for repair order to any customer home. It also allows easy resupply of GWCs and retail stores, which receive replenishment shipments frequently. All shipments are sent through small parcel carriers such as FedEx, UPS, or USPS.

In addition to forward logistics, there is a clean-and-screen facility co-located with one of the RDCs and run by a separate third party service provider. This facility is responsible for receiving used equipment from end users, inspecting it for damage, repairing it, testing it to ensure functionality, and cleaning it. If a device is broken and unrepairable it is returned to the manufacturer (for warrantied items) or scrapped. Once a functional device is tested and cleaned, it is placed back in RDC inventory for reuse with a new customer. This practice allows Verizon to limit its new equipment purchasing costs while ensuring that the customer receives quality equipment.

As Section 2.1 will discuss, certain aspects of the Fios supply chain are common across a wide variety of industries. In particular, clean and screen and reverse logistics have become ubiquitous in the modern economy. The rise of online shopping has required all manner of businesses to process returned goods and return them to use when possible. The forward supply chain of Fios, as characterized by centralized RDCs and varied local distribution channels, is also common across industries. Overall, many problems inherent in the Fios supply chain are frequently seen across a broad spectrum of business, particularly those operating in service-focused industries.

1.4 Problem Statement

Every supply chain suffers from inefficiencies and waste. This thesis conveys the results of efforts to reduce waste associated with the shipment of Verizon Fios customer premise equipment (CPE). CPE, such as set top boxes, broadband routers, and WiFi extenders, is shipped to Fios customers via the three channels previously mentioned. Each method has its own benefits and shortcomings due to impacts on metrics such as inventory, shipping costs, on-time delivery, and system complexity. Although the majority of shipments are successfully activated in the customer's home, a non-trivial percentage results in unused returns or inventory shrinkage. These undesirable results represent a significant amount of wasted resources. This project analyzes the system and presents a framework to find cost savings in the Fios supply chain by reducing the waste resulting from such unsuccessful shipments. It also presents findings related to technology employed within the supply chain such as RFID that may also relate to inefficiencies and waste.

1.5 Research Methodology

Identifying the key differences between how a system currently works and how it is ideally meant to function will highlight critical savings opportunities. Thus, the first phase of this project involves building a basic understanding of the current Fios business, its underlying processes, and the data that drives decision-making. Through a literature review, interviews with key stakeholders across the supply chain, data collection from the various key systems, and analysis of shipment data, this phase identifies particularly strong targets for savings opportunities. Once these key features are identified, the project pursues a two-pronged approach. This involves focusing work on "internal" process analyses to identify shortcomings in Fios internal business practices and systems, alongside "external" analysis to control for factors outside of the system. External analysis leads to predictive models capable of limiting waste through prevention of shipments with an elevated risk of customer cancellation.

1.6 Thesis Structure

This thesis is organized into six chapters. The content of each chapter is summarized below.

- Chapter 1: This chapter provides an introduction to Verizon, Fios, and the Fios supply chain. In addition, this chapter includes a brief problem statement, description of research methodology, and thesis structure.

- Chapter 2: A brief literature review summarizes previous academic research work that relates to efficiency within closed loop supply chains, device tracking technologies, and consumer behavior predictions
- Chapter 3: This chapter focuses on the Fios supply chain's operations, organizational processes, data collection systems, and key data which produced initial insights. This chapter also includes analysis of device tracking technologies employed within the supply chain. Finally, it details the techniques used to analyze the current system via process mining techniques.
- Chapter 4: The formulation and results of predictive modelling efforts are presented. The limitations of the model are highlighted, as well as the proposed applications.
- Chapter 5: The conclusion of the thesis includes recommendations on the application of predictive modelling within closed loop supply chains. In addition, specific recommendations for Verizon Fios supply chain are presented. Opportunities for future research in process mining, organizational dynamics, and predictive modelling are identified.

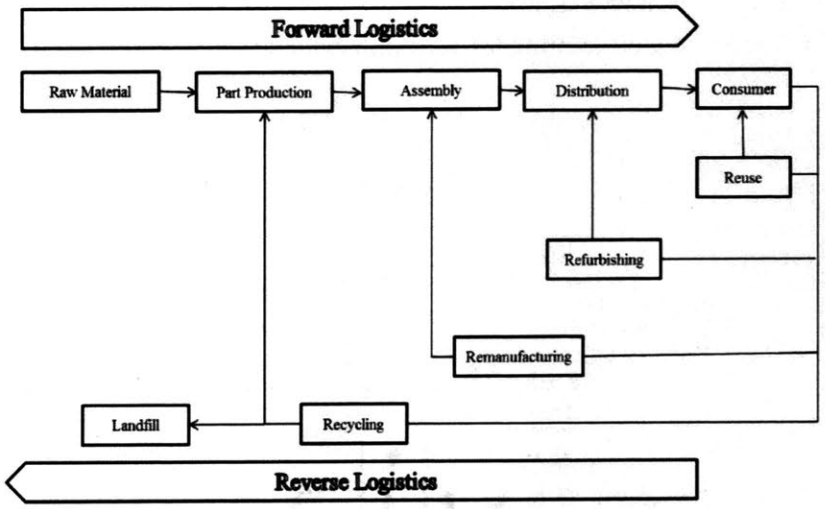
2 Literature Review

Companies across the world are increasingly incorporating reverse logistics and closed loop supply chains into their operations. These arrangements, when feasible, can have positive impacts on supply chain costs, efficiency, and environmental impact. In addition, data collection and supply chain tracking technology has made significant advances in recent decades, allowing unprecedented visibility of the movement of goods, as well as the advent of "big data" and analytics within supply chains. This chapter reviews recent research and market trends surrounding these topics.

2.1 Closed Loop Supply Chains and Reverse Logistics: Overview

Reverse logistics is defined as ''The process of planning, implementing, and controlling the efficient, cost effective flow of raw materials, in-process inventory, finished goods and related information from the **point of consumption** to the **point of origin** for the purpose of recapturing value or proper disposal'' [4]. When a reverse logistics mechanism is combined with a product's standard forward logistics system, a closed loop supply chain (CLSC) is formed. Among other things, a proper CLSC requires establishment of mechanisms for retrieval, refurbishment, remanufacturing, and recycling or disposal of used equipment. Figure 2-1 describes the major components of a typical CLSC [6].

Figure 2-1: Major Components and Activities of a CLSC



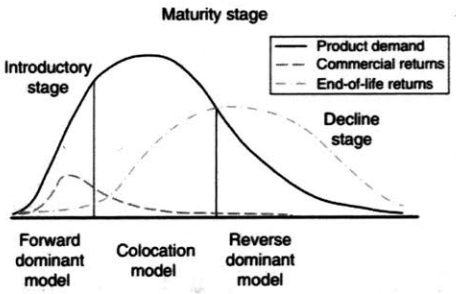
For many companies, the initiation of closed loop supply chains was the result of legal directives mandating collection and recovery efforts for certain types of waste. However, in recent years, numerous industries have come to understand the potential for value creation in CLSCs. Toffel lists the following as some of the prime motives for companies to pursue product recovery through reverse logistics [5]:

- Reducing production costs
- Promoting an image of environmental sustainability
- Meeting customer demands
- Protecting aftermarkets
- Preempting regulation

Although the article was written through the lens of product recovery in manufacturing, these same motives apply to the Fios supply chain. For example, "meeting customer demands" may refer to the desire of most Fios customers to rent their equipment in lieu of outright purchases. This arrangement necessitates a mechanism to receive used goods at the end of the equipment's life cycle (or the customer's tenure), and thus a CLSC is an important component in the Fios business model.

There are numerous challenges to operating a CLSC, and one of the most commonly discussed in literature is the retrieval of equipment. Sahyouni notes that there are typically two main types of return: commercial and end-of-life. Commercial returns are often a result of customer preferences or product defects. End-of-life returns are the result of product life cycles and new product introduction. Figure 2-2 displays the typical timing distribution of these return types. In the same study it is noted that, due to inadequate incorporation of reverse logistics into supply chain planning, "companies now face a considerable challenge in designing a reverse supply chain network that will meet returns processing needs while complementing their existing forward distribution system" [7].

Figure 2-2: CLSC Return Shipment Timing



A Deloitte strategic review also highlights difficulties of equipment retrieval, particularly in the predictability of return volumes. It argues: "reverse logistics happens in response to an action of a customer or supply chain actor and as such is extremely difficult to anticipate or plan for by a company" [8]. Verizon mitigates some of the return volume risk by employing a 3PL provider to process, test, and

refurbish its returns. Under this agreement, Verizon pays per unit processed and is therefore insulated from daily shipment fluctuations in volume which may create short-term labor shortages or overages.

Despite these challenges, the importance of reverse logistics continues to grow around the world. By Deloitte's estimate, reverse logistics has annual cost in the U.S. of about 200 billion USD. In order to minimize these costs and ensure successful operation of CLSCs, Deloitte recommends the following as key factors for success:

- Optimize forward logistics – Minimize customer returns by implementing the correct strategy in forward logistics to limit impacts on the reverse flow
- Synergies – Merge forward and reserve flows
- Product return policy – Product return policies should not only be looked at from a commercial perspective though should be considered from a logistics and operational point of view as well.
- Consolidation of flows – The success of a reverse flow depends on the degree of convergence between the financial flow, operational flow as well as the information flow [8].

Verizon has addressed synergies by collocating its reverse and forward logistics facilities and it continuously works to optimize its forward logistics to mitigate waste. The Fios supply chain generates data during each transaction and movement of goods within the system. Thus, there are significant opportunities to mitigate waste through analysis of processes, technologies, and big data analytics.

2.2 Big Data and Predictive Analytics in Supply Chains

Due to its somewhat nascent nature, "big data" lacks universal definition. This thesis will use Gartner's definition, which describes big data as "high-volume, high-velocity and/or high-variety information assets that enable enhanced insight, decision-making, and process automation" [9]. Using big data, researchers generally approach problems using three categories of analytic method. Table 2-1 defines the methods and provides brief examples [9]. At first glance, it may appear that prescriptive and predictive analytics serve similar purposes since they are both forward looking. There is, however, a key distinction. Predictive analytics merely describe a potential future outcome, whereas prescriptive techniques recommend specific actions and generally seek to understand the predicted impact of these actions on performance.

Table 2-1: Analytics Methods and Examples

Analytics Method	Definition	Use Case Examples
Descriptive	Used to describe or mimic the system or process under study and answer the question of what is happening	• Supply chain mapping • Model risk analysis • Model supply chain flexibility
Prescriptive	Used to prescribe to the decision-maker some "optimal" set of policies and answers the question of what will be happening	• Production planning • Project selection • Profit Maximization • Vehicle routing
Predictive	Provide a projection of system or process performance into the future and answer the question of what will be happening	• Demand forecasting • Returns predictions

This thesis is focused primarily on predictive analytics within reverse logistics, specifically predicting customer cancellations. Literature on this specific topic is somewhat sparse, but one particular study by Deshmukh outlines some of the key challenges with such an undertaking [11]. The first of these is model selection. Various machine learning and optimization programs are feasible in this use case, but Table 2-2 outlines a comparative study between different techniques across several features [11]. The key takeaway from this table is that it is critical to understand the conditions under which a certain technique outperforms others for a given problem.

Table 2-2: Comparing learning algorithms (ranked from * to ** (best model))**

	Machine Learning Techniques				
Comparison Metrics	Decision Trees	Neural Networks	Naive Bayes	kNN	Support Vector Machines
Accuracy in general	**	***	*	**	****
Speed of learning with respect to number of attributes and the number of instances	***	*	****	****	*
Speed of classification	****	****	****	*	****
Tolerance to missing values	***	*	****	*	**
Tolerance to irrelavant attributes	***	*	**	**	****
Tolerance to redundant attributes	**	**	*	**	***
Tolerance to highly interdependent attributes (eg. parity problems)	**	***	*	*	***
Dealing with discrete/binary/continuous problems	****	*** (not discrete)	*** (not continuous)	*** (not directly discrete)	** (not discrete)
Tolerance to noise	**	**	***	*	**
Dealing with danger of overfitting	**	*	***	***	**
Attempts for incremental learning	**	***	****	****	**
Explanation ability/transperancy of knowledge/classifications	****	*	****	**	*
Model parameter handling	***	*	****	***	*

A second challenge in predicting customer cancellations is that the data is typically imbalanced, with far fewer cancellations than completed orders. In order to mitigate the effects of this imbalance, two techniques are proposed: (1) down-sampling the majority class or over-sampling the minority class or both, and (2) cost-sensitive learning, i.e., assigning a high cost to misclassification [11]. In addition, imbalance-treated random forests and gradient boosted classifiers are identified as high performing techniques for this type of data set.

Finally, a third challenge outlined by Deshmukh is the choice of evaluation metrics. It is noted that overall model accuracy is not appropriate to evaluate highly imbalanced data sets. Using this metric, a model built using a dataset containing 1% cancellations would be viewed as 99% accurate if it simply predicted zero cancellations. In this case, sensitivity and precision are more appropriate performance metrics. Table 2-3 shows the formula and interpretation for each evaluation metric [11].

Table 2-3: Formulae and interpretation of accuracy, precision and recall scores

Evaluation Metric	Formula	Interpretation
Overall Accuracy	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$	This metric says how often is the classifier correct
Sensitivity / Recall	$\frac{(TP)}{(TP+FN)}$	When an instance actually falls within a class, how often does the model correctly classify it as falling in this class
Positive Prediction Value (PPV) / Precision	$\frac{(TP)}{(TP+FP)}$	When the model predicts an instance to fall within a class, how often does it actually fall within the class

In addition to these challenges, another issue that confronts predictive modelling is the idea of concept drift. According to this idea subtle changes in processes over time alter the underlying probability distributions of the data. The effect is that learning models trained on old data may be inconsistent with new data [15]. In order to mitigate the effect of this, Chen proposes implementation of incremental learning, whereby the data stream is divided into chunks and classifiers are trained on each data chunk and combined to predict outcomes in the newest data chunk [15].

Although the prospects of building strong predictive models are good, Hazen provides an insightful warning: "The modern world is data rich and thus big data analytics is likely here to say. However, management decisions are only as good as the data on which they are based" [9]. Thus, any organization which purports to make use of big data analytics must first lay the groundwork of reliable IT infrastructure and data collection methods. In addition, Hazen describes socialization of analytics as another major challenge. He argues that transitioning an organization to rely on big data analytics requires careful strategic planning and careful consideration of the cultural and organizational ramifications of such a tectonic shift in business focus. Although Verizon has laid a strong foundation for incorporation of big data analytics into supply chain decision-making, this thesis outlines some impacts of problems with consistency of data that may inhibit the effectiveness of predictive modelling or its implementation.

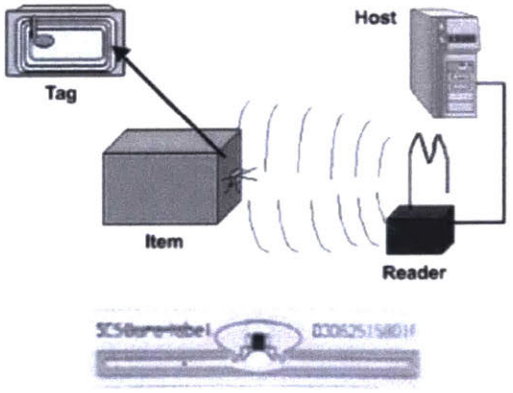
2.3 Radio Frequency Identification Implementation and Verification

A major factor in the advent of big data is the proliferation of cost-effective data collection and device tracking technologies. Radio frequency identification (RFID) tags are at the forefront of this revolution. These devices were first introduced commercially in the 1980s via active, or battery powered, tags. These tags were prohibitively expensive and generally used only for the tracking and management of extremely valuable property. The more recent introduction of passive RFID technology has drastically

reduced the cost and increased the accessibility of RFID tags across the economy [12]. Today, Statista estimates global annual RFID sales to be about 17 Billion USD, with double digit annual growth rates over the next decade [13].

Figure 2-3 depicts key RFID system components and functionality. Each RFID tag is embedded with a microchip that contains identification data such as an Electronic Product Code (EPC). The microchip can also incorporate functionality beyond simple identification, including integrated sensors, read/write storage, encryption and access control. The tag is typically attached to an item, case or pallet, and the enterprise RFID database is updated to reflect the EPC/serial number pairing. RFID scans are conducted with readers. These devices contain small antennae that emit electromagnetic waves which form a magnetic field when they "couple" with antenna on the RFID tag. The tag draws power from the magnetic field and uses it to power the microchips' circuits. The microchip then modulates the received signal in accordance with its identification or programmed code and transmits or reflects a radio frequency signal. The modulation is in turn picked up by the reader, which decodes the information contained in the transponder and depending upon the reader configuration, either stores the information, acts upon it, or transmits the information to the host computer and database [12].

Figure 2-3: RFID Components and Functionality



RFID Tag (not to scale)

The Fios supply chain is equipped with EPC-encoded passive RFID tags on the vast majority of its customer premise equipment. Each tag contains company identification, an item description, and item serial number. The organization employs a mix of bar code scanning and RFID scans. Typically, palletized or boxed equipment at centralized facilities is identified using bar code scans for shipments, and RFID scans are used for individual item inventory at more remote locations.

The widespread implementation of RFID has numerous supply chain benefits. Attaran identifies a few such benefits when compared to existing device tracking technologies [22]:

- Enhanced visibility along the supply chain
- Speedy and accurate information retrieval
- Accurate asset tracking
- Better-quality information
- Improved productivity
- Reduced operating costs
- Improved business process
- Improved quality & reliability

He also notes that RFID tags are capable of storing far more information than bar codes. In addition, the ability to scan multiple RFID tags simultaneously and without direct line of sight is a drastic improvement over often-tedious barcode scanning technology. These improvements allow the potential for supply chain automation on an unprecedented scale, as it is possible to establish scanning protocols that allow real-time updates of inventory and goods movement all across the supply chain, from order fulfillment to reverse logistics.

The widespread implementation of RFID in supply chains is not without challenges. RFID tags are more expensive than simple barcodes and it is often difficult to articulate a positive return on investment in the technology. In addition, RFID is a wireless technology and, as such, poses some potential security concerns to users regarding the compromise of data during wireless transmission, storage of data, and security of storage sites [13]. It is theoretically possible for an RFID tag with read/write capability to have its EPC information altered by nefarious third parties, or for competitors to glean information on the movement of goods via RFID interception. Some of the security issues have been addressed by RFID vendors by employing varying querying protocols, jamming, encryption, and other techniques. Despite these challenges, RFID technology is capable of providing a significant boost to supply chain productivity and visibility. The Fios supply chain employs RFID on a broad scale, and plans are in place to broaden usage of the technology to allow higher granularity data generation. As this thesis will discuss, there are significant improvements that should be made to RFID information flows before such widespread change should occur.

3 Understanding Current Operations: Modelling Existing Systems

This chapter discusses the results of initial system analyses. It discusses the organizational structure and nature of the existing system and the collection of data from a variety of disparate sources. This chapter also includes analysis of device tracking technologies within the Fios supply chain, and presents a key opportunity for process improvement related to those technologies. Finally, process mining is discussed as a means to understand event sequences and system interactions.

3.1 Fios Supply Chain Organization and Challenges

The first step in solving any significant problem is understanding the underlying systems and structures. A key component of the supply chain are its constituent stakeholders. Table 3-1 outlines some of these key stakeholders in the movement and delivery of products, as well as their primary considerations as defined by their performance metrics.

Table 3-1: Key Stakeholders and Motivations

Stakeholder	Primary Considerations
3PL Providers	Item processing volume
Logistics Services	Low Inventory holding and procurement costs
Sales and Marketing	New customer attraction and retention
Customer Service	Expedient issue resolution
GWCs	Expedient installation and repair

Across the supply chain, there is a wide disparity of motivations and goals between various organizations. These goals, rooted in individual and organizational performance metrics, are often highly logical for the specific organization but when viewed holistically they can be seen to drive the overall business away from a global optimal solution. A strong example of this can be seen in customer service. One the most important performance metrics for customer service representatives is short call duration. This is a logical metric designed to increase the efficiency of the company's call center employees (it should be noted that other metrics are in place to ensure a positive customer experience). One of the easiest ways to potentially please a customer and quickly end a call is to ship replacement equipment to the customer's home overnight. Therefore, a customer service representative may choose to drop-ship after pursuing only the simplest of remedies.

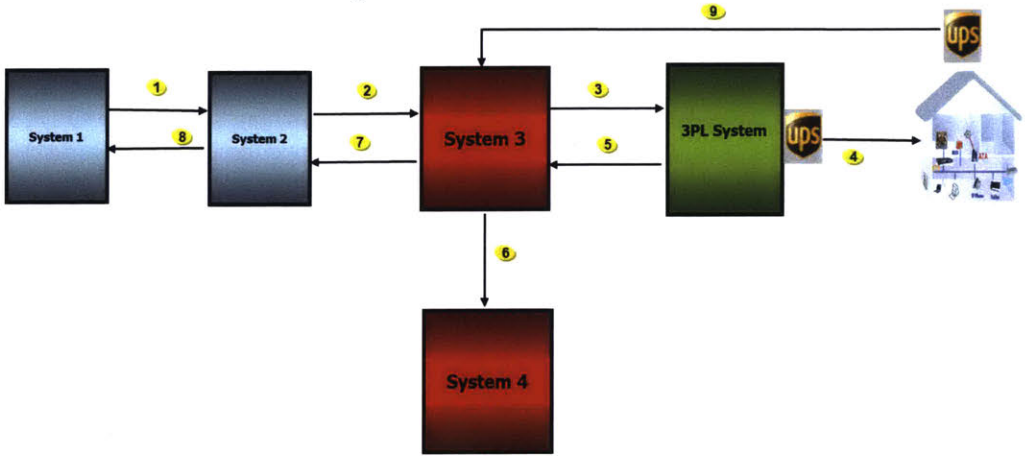
A similar example can be seen in the performance of repair technicians. They are highly motivated to complete each job as quickly as possible. As a result, there is incentive to simply replace potentially faulty equipment instead of initially committing to in-depth troubleshooting. Clean and screen

testing results indicate that a vast majority of returned devices are fully functional. Thus, a system which incentivizes short duration technician visits may actually result in the unnecessary use of additional equipment. These examples depict possible sources of inefficiencies within the holistic CLSC, but a topic of further research may seek to understand the tradeoffs between constituent organizational incentives and the impact on the overall business.

3.2 Data Sources and Pitfalls

In addition to the stakeholders, it is also critical to understand the IT infrastructure within the supply chain. As mentioned in Section 1-1, Verizon is the child of numerous mergers and acquisitions. Constant flux and steady growth throughout the company's history have resulted in a wide range of data systems (each with its own unique set of success metrics) which drive the Fios supply chain. Figure 3-1 maps information flow for a hypothetical direct ship order.

Figure 3-1: Order Information Flow



In this flow, an order is initially placed through a sales processing system, which sends information to the customer premise equipment system, resulting in an order placement in Verizon's logistics software. This program then coordinates with the 3PL provider's system to initiate the shipment. Notification of the completed shipment initiates system updates in the reverse order, and the logistics tracking software also updates inventory records in the company's enterprise resource planning software. Outside of this shipment flow, there is also a system which oversees the provisioning of internet and TV services to the customer which must be updated with correct equipment shipment information. Failure to do so can result in equipment installation or service activation difficulties.

In short, the successful completion and tracking of a shipment from Fios may involve no less than six IT systems. In an ideal world with perfect information transfer this complex architecture would not be

a problem. Unfortunately, analysis of shipment data from each system shows non-insignificant discrepancies between systems. As a result of these data transfer issues, it can be difficult to determine which data sets are accurate for specific features. Table 3-2 outlines some of the common apparent data issues within the primary systems, in addition to the system-to-system mismatches.

Table 3-2: Data Concerns by Source

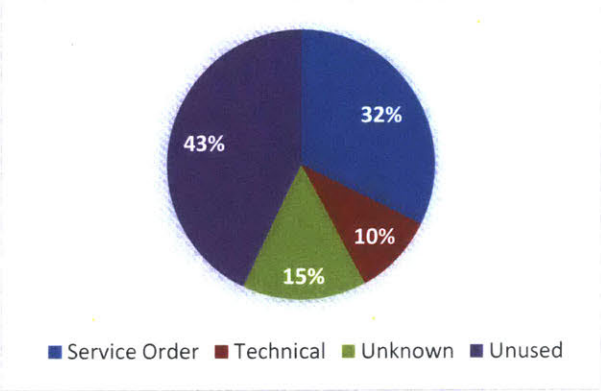
System Description	Common Problems
Clean and Screen / 3PL	• Missing receipts (multiple shipments of devices with no intervening receipt) • Item received same day as shipped from RDC
Logistics Tracking	• Due date mismatches • Order status discrepancies
Customer Premise	• Simultaneous activations/deactivations
Provisioning	• Missing aggregate repair order data

Ultimately, it is clear that no single source of data within this system provides sufficient accurate information to conduct analysis with any assurance of veracity. It is necessary to produce combined data sets, focused on a particular type of equipment, and cherry-pick the most accurate features from each constituent data system. Using these combined data sets, it is possible to garner insights to guide specific areas to research further in the quest to eliminate waste.

3.3 Characterization of Existing Shipment Results

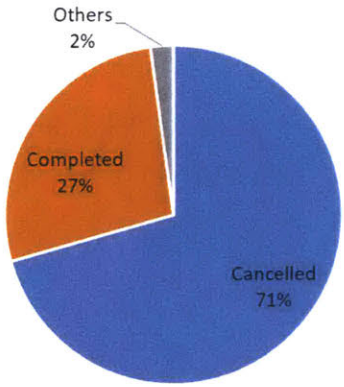
Although neither complex nor challenging, basic analysis of the aforementioned data is worth discussing in brief, simply to convey the scale of the problem and motivations for further lines of effort. Figure 3-2 shows the prevalence of unused returns at clean and screen, by type. These devices arrived back at the RDC having never been activated on the network, and typically still in the original packaging. They are of particular concern because they represent pure waste, in terms of shipment, 3PL processing, and inventory holding costs.

Figure 3-2: Clean and Screen Returns by Type



This result shows the existence of major barriers to efficiency within the CLSC, but further analysis demonstrates that direct shipments on technician installation orders are the main source of unused returns. In fact, new technician installs account for about two-thirds of direct shipments to customers, but over 85% of unused return volume. Even further, Figure 3-3 shows that cancelled orders are the primary driver of unused returns.

Figure 3-3: Unused Return Order Status



It is important to note that the reason Verizon ships equipment for new technician install orders is to limit inventory levels at its hundreds of GWCs. Technicians frequently draw equipment from GWC inventory for installation and repair orders, but the direct shipment program is simply another option that allows Verizon to centrally hold inventory at the RDC to limit the effects of demand variability. Thus, the direct shipment program is not entirely wrought with waste, and does serve to lower Verizon's supply chain costs. It can, however, be improved.

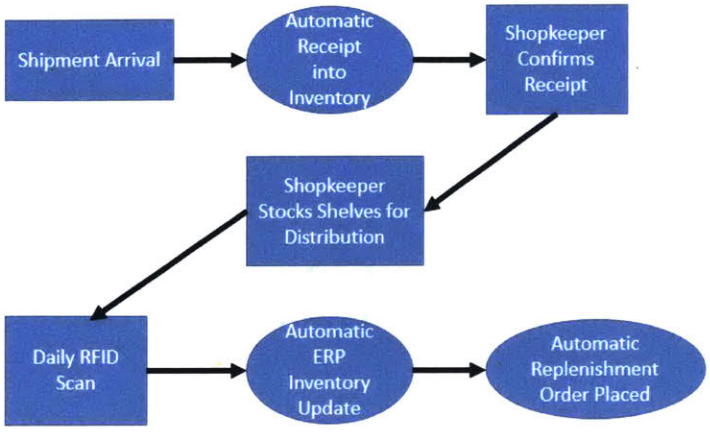
From these initial insights and the preceding infrastructure mapping, two lines of effort are pursued. First, it is obvious that increased understanding of internal processes and systems is required. Thus, device tracking systems, particularly RFID, are investigated to examine existing usage and opportunities for future employment. In addition, process mining is pursued as a potential means to increase understanding of process flows and identify information barriers. Second, predictive efforts to limit shipments resulting in cancellations resulted in a random forest model to be discussed in Chapter 4.

3.4 Analysis of Device Tracking Technologies and Current Usage

As discussed in Chapter 2, RFID usage has become a significant asset for supply chain managers across the globe, and Verizon is no different. The company currently employs RFID on the majority of its customer premise equipment to provide asset visibility from the RDC all the way to device installation. These RFIDs are installed directly onto the equipment by the OEM, and the tag information is transferred into Verizon's database at or around the time of physical receipt at the RDC. Bulk tracking and transactions at the RDC are generally conducted via barcode scanning; palletized and boxed equipment generally has consolidated barcodes on the outside and the system recognizes that a scan for one device is a scan for all devices in the container. RFID becomes a critical component further down the supply chain to track movement of individual devices.

In order to investigate the use and effectiveness of RFID for inventory tracking, we visited a GWC to understand the process and discuss any difficulties with the end user: the GWC "store keepers". These individuals are entrusted with accountability for the GWC's equipment and conduct RFID-based inventories on a daily basis. Each garage has two inventory categories: "T" stock and "S" stock. T stock, or "truck stock" refers to devices ready to move or already on a truck for delivery to a customer premise. S-stock denotes devices that are in GWC inventory that have not been made ready for use. This distinction is critical because replenishment orders are placed automatically based on S-stock levels. Figure 3-4 shows the flow of equipment and information through the GWC.

Figure 3-4: Inventory Process Flow



In this process, equipment arrives early in the morning, typically prior to the arrival of any GWC personnel. Devices are automatically received into GWC inventory S-stock when the delivery driver scans the pallet barcode. Store keepers bar code scan the equipment to confirm the morning's automatic receipt and place the equipment on shelves for use. Later in the morning, after the garage's technicians have departed to serve customers, the store keepers conduct an RFID scan to inventory the garage's equipment. Any RFID-labelled equipment that is not detected in this scan is automatically placed into T-stock as it is assumed to be with a technician. If this inventory brings the GWC S-stock below a certain threshold level, a replenishment order is automatically generated for equipment to arrive the following morning.

This system is extremely efficient when everything is running smoothly. It typically allows a storekeeper to conduct daily inventories of hundreds of items in less than 15 minutes. However, RFID usage in this garage provides an excellent case study in the value of the "Gemba" Lean concept, which alludes to visiting the place where value is created. In this case, the GWC's storekeepers informed us and demonstrated that numerous devices are equipped with readable tags that are unidentifiable to Verizon's ERP system. In one particular example, we observed an inventory scan of a pallet containing 80 devices. The initial scan of these devices detected 80 RFID tags, as expected. However, when this list of tags was transferred from the scanner into the ERP system, a small number of tags were rejected as invalid with no associated serial number. As a result, the ERP system was automatically updated to reflect an S-stock inventory count that was reduced by the number of rejected tags. This result was duplicated numerous times and according to the store keeper had been a recurring problem for quite some time, particularly with used devices coming from the clean and screen facility. This anecdotal evidence is immediately

verifiable with available data. A comparison of the RFID tagged assets database with the ERP's inventory database shows some mismatches and some missing data. One random sample of over 1000 serial numbers showed that greater than 5% of devices in the ERP inventory were either not present in the RFID tagged assets database, or had a mismatched serial number.

Although root causes of the problem are still under investigation, this episode highlights a key issue with RFID that is not discussed at length in the literature: use of RFID, particularly in conjunction with 3PL providers, adds complexity to a company's IT systems and careful consideration must be made to ensure the fidelity of data. In this case, early signs indicate that data transfer between 3PL providers and Verizon is sometimes incomplete. When a device returns to the clean and screen facility and the RFID tag is unreadable, a replacement tag is issued and programmed to match the device. The tag identification information should then be transferred to Verizon to allow an update within the ERP database. It appears that this transfer of data from the 3PL to Verizon is not always occurring on a consistent basis.

It is difficult to determine the immediate impact of this problem on the supply chain due to the "snapshot" nature of the relevant data (the database only displays the current information for a device and lacks transaction history). It is likely, however, that such missing data has resulted in inefficient operations in the past and will continue to do so unless rectified. The main source of this inefficiency lies in the daily inventory's automatic updates of on-hand inventory and subsequent automatically triggered replenishment orders. In a hypothetical scenario in which Verizon lacks the RFID EPC information on just 10% of refurbished devices, automatic replenishment orders will frequently result in premature shipments of equipment (before the physical inventory count actually reaches the target threshold). As a result, Verizon incurs significant additional inventory costs and many key supply chain planning factors such as days of supply become skewed. As mentioned in Section 2-2, management decisions are only as good as the data on which they are based.

3.5 Process Mining to Understand System Interactions and Trends

Sections 3-1 and 3-2 show examples of challenges associated with complex organizational and IT structures. As a result of these types of challenges, there is some uncertainty for key decision makers regarding both data and process fidelity. In order to better understand this uncertainty, this research seeks to map true process flows of both information and inventory in order to ascertain where reality may be deviating from the published, idealized processes. The goal is twofold: (1) to establish an updated realistic

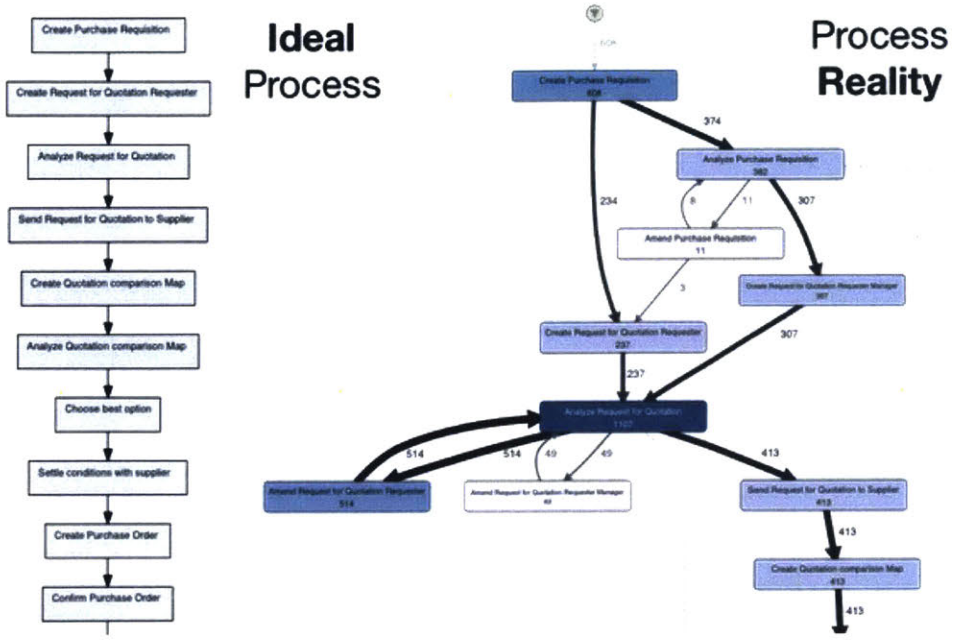
process flow, and (2) to use process mining to ascertain whether any specific deviations from the "normal" pathways are resulting in wasteful shipments or returns.

3.5.1 Process Mining Description

Process mining is a relatively new concept that refers generally to techniques that allow users to extract data from event logs. Researchers claim that it serves as the missing link between model-based process analysis and data-oriented analysis techniques, allowing users to perform fact based business process management [14]. The starting point for process mining is the aforementioned event log. Each event in such a log refers to an activity (i.e., a well-defined step in some process) and is related to a particular case (i.e., a process instance). The events belonging to a case are ordered and can be seen as one "run" or "trace" of the process. Event logs may store additional information about events such as the resource (i.e., person or device) executing or initiating the activity, the timestamp of the event, or data elements recorded with the event (e.g., the size of an order) [14].

There are three main types of process mining outlined in the literature. The first type of process mining is discovery. Process discovery is the most prominent process mining technique and it involves taking an event log and producing a model without any prior knowledge of the process. The second type of process mining is conformance. Here, an existing process model is compared with an event log of the same process. Conformance checking can be used to check if reality, as recorded in the log, conforms to the model and vice versa. Figure 3-5 depicts an example a process map from process mining compared to its idealized model.

Figure 3-5: Conformance Checking in Process Mining



The third type of process mining is enhancement. Here, the idea is to extend or improve an existing process model by using information about the actual process recorded in some event log. Whereas conformance checking measures the alignment between model and reality, this third type of process mining aims at changing or extending the a priori model. For instance, by using timestamps in the event log one can extend the model to show bottlenecks, service levels, and throughput times [14].

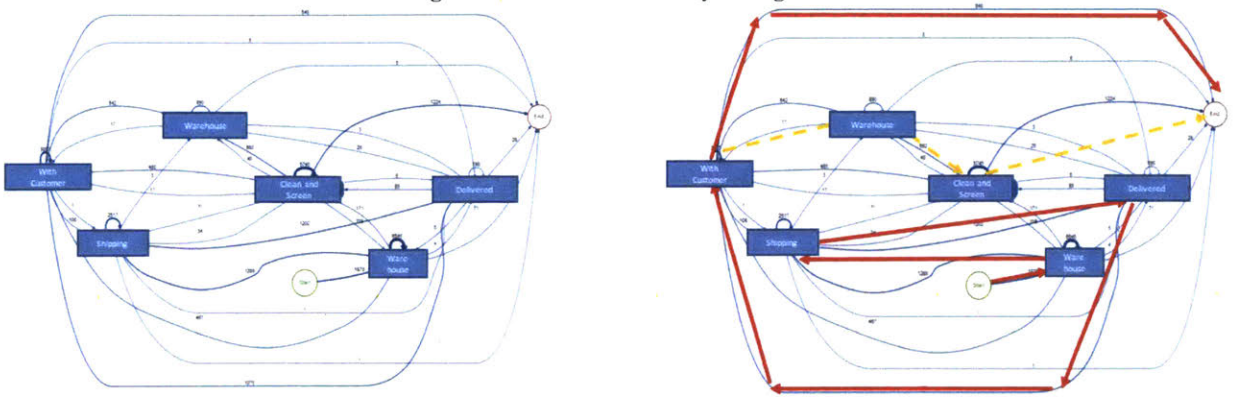
There are a number of both commercial and open source tools with which to conduct analysis. In order to avoid incurring costs for this exploratory work, we chose to work with a pair of open source tools: the 'R' package bupaR, and the open source software ProM Tools. BupaR is significantly more user friendly and produces far more aesthetically pleasing outputs, but ProM Tools is far more versatile for the expert user.

3.5.2 Application and Results

Within the Verizon IT infrastructure, event log data is sparse. To limit file storage sizes, many databases store only the snapshot of device activity, depicting only the most recent transaction. There are, however, a few systems that store such event data in an aggregated format that proved easy to extract. One such system is the logistics management software that Verizon uses. The event logs from this database allow the researcher to trace the movement of a device from the RDC shelf to the customer's

house and back to clean and screen. Figure 3-6a depicts the process flow for the event logs within this system, and 3-6b highlights the main transactions that occur.

Figure 3-6: Process Discovery in Logistics



(a) (b)

The key takeaway from Figure 3-6 is that, while the main process flow is discernable based on volume along that path, there are significant deviations from the normal process flow that occur on a regular basis. These pathways are sometimes entirely illogical, such as a device moving from "delivered to customer" straight to "in stock at RDC." This may be accurate, and devices may skip steps in the process at certain times, but more likely it is an indicator of poor data (either from poor collection or missing transactions). Efforts to expand the mapping by combining multiple systems are severely bogged down by these "outlier" transactions. The maps become largely unreadable, and the dominant flow path is difficult to discern.

After generating initial process maps, we endeavor to compare the processes that result in 'negative,' or wasteful outcomes, against those with positive outcomes. In this case, a positive outcome means an activation at a customer's premise, whereas a negative outcome means any other circumstance, e.g., unused return to clean and screen, delivered but never activated, etc. These comparisons show that there is significant process variability across the entire population, but the 'bad' population is far more random in its transaction sequences. Out of hundreds of bad shipments, the most common 'trace' sequence only accounts for 2% of the population. The good shipments were slightly less variable, although not nearly as regular as one would expect given that those shipments were all ultimately activated and theoretically should have followed a similar flow of transactions. The most common 'trace' sequence for good shipments accounts for about 15% of the population.

Process mining shows significant potential to yield key insights into process flows. It is relatively simple to conduct discovery process mining to create a new process map. Using process mining techniques on data from the individual Fios systems, we are able to map out information and equipment flows, and identify the dominant paths within that system. Complications arise, however, when we attempt to integrate data from multiple systems. Discrepancies between systems present a significant data validation challenge and varied data formats between systems prove difficult to integrate. As a result of these challenges, we are currently unable to integrate and map process flows across the entire Fios supply chain with any sort of certainty. This may become possible in the future if the company seeks to unify or integrate its disparate systems into a more comprehensive solution.

4 Predictive Modelling for Order Cancellations

This chapter describes the methods and techniques used to build a predictive model that reliably predicts customer cancellations, as well as the key considerations that are required. Section 4-1 describes the motivation behind such efforts and briefly describes the data sources and evaluation criteria used to select the model.

4.1 Method and Motivations

This section outlines the reasoning behind the predictive model and some of its key considerations. In addition, it discusses data sources and methods employed.

4.1.1 Motivation and Goals for Predictive Model

In Section 3.4 we outline the results of a simple Fios shipment outcome analysis. The key opportunity for predictive analytics is identified as shipments on cancelled orders, which overwhelmingly result in unused returns or missing equipment. These lost items and unused returns cost Verizon millions of dollars per year in direct device acquisition costs, clean and screen costs, and packaging and shipment costs. Recognizing that there is at least some inherent random nature in the behavior of customers, the stated goal of this model is to produce actionable insights that can limit the waste resulting from these shipments. Some of the considerations for the construction of this model include:

- Implementability: the model must be simple to disseminate and provide real-time predictions on orders. If used in a prescriptive manner to make shipment decisions, the model must be incorporated into the workflow of sales associates, and so it must be quick, user-friendly, and clear in its result.
- Accessibility of data: the model must be built using data that is available at the time of order placement. Although a plethora of data is available for a rearward looking researcher, model features are not useful for predictive modelling unless they are actually predictive.
- Interpretability and communicability: In order for management to make decisions, they must first understand the model and its results. Clear methods to visualize models are imperative in business settings.
- Impact on inventory: the primary reason for the existence of the direct shipment program is to limit GWC inventory. False positives (erroneous cancellation predictions) would result in equipment being installed from GWC inventory that would have otherwise been shipped. It is therefore imperative to limit false positives in the final model.

4.1.2 Methods: Data Sources and Treatment

In order to limit the effects of concept drift, we choose to limit data samples to within the past year. The simplest way to build a readily implementable model is to build it entirely with data from the order processing database. For this model, each direct shipment order is associated with over 300 features from the order processing system. After cleaning the features that are either entirely null or entirely uniform, we are left with 50 distinct features from the order processing system. For initial model building, we select a training sample from over 100,000 device shipments over a six-month period. Some of the key features in the order system include:

- Order placement and due dates
- Account establishment date
- Existing Internet/TV service type
- Residence/Address Type (single family home, multi-unit apartment, small business, etc.)
- Sales Agency: Verizon employs sales representatives in-house, and also outsources sales to third party referral agencies
- Service type: Verizon offers bundle packages of any combination of internet, phone, and television services.

For the dates, a difference between each feature is calculated in days and employed as a feature in the model. For example, the count of days between order date and due date is listed as the feature 'Order vs Due Date'.

In addition to the order processing system, Verizon also uses other, more comprehensive, data sets within its sales organization. This data is generally more rich in terms of features and content than the ordering system data, but can be more difficult to access and integrate into real time systems. The specific features are proprietary. A second model is developed and tuned with this data set (henceforth known as sales data), and the expectation is that this model will be more accurate than the original, but also more difficult to implement as it requires live input from multiple systems.

For each data source, there are significant impurities within the sample data. Rather than impute the missing values with a calculated value such as the mean of the feature, we instead choose to recognize missing data as its own unique value, understanding that the very fact that data is missing may be a clue in predicting the order outcome. For categorical features, these blanks are managed through one-hot encoding. In this process, categorical variables are converted into a new set of binary features, with one new feature for each unique categorical value. For example, a categorical feature 'AB' with unique

variables 'A', 'B', and 'Blank' will be converted into three new features: AB_A, AB_B, and AB_Blank. Each of these new features is binary with, for example, a value of 1 in AB_A for each instance of 'A' in the original categorical variable. One-hot encoding is important because it allows for the employment of machine learning algorithms that are otherwise incapable of handling categorical variables, such as Random Forest. For the numerical features, missing values are replaced with the value of 1,000,000. This arbitrary number is chosen to ensure a value well outside the range of any of the actual feature data. Thus, a random forest classifier will easily distinguish the real data from the inserted values.

4.1.3 Methods: Evaluation Criteria

As mentioned in the literature review, overall accuracy score is a poor metric to evaluate models with highly imbalanced data. Therefore, we seek alternative methods that will allow us to evaluate a model based on its ability to both detect true positive and limit false positives. In addition to the recall and precision metrics outlined earlier, we also use the false positive rate, defined as:

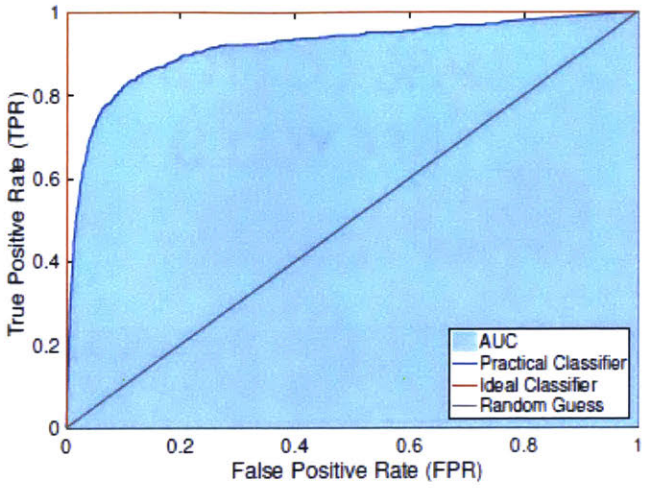
Equation 4-1: False Positive Rate

$$FPR = \frac{FP}{FP + TN}$$

where *FP* represents the number of false positives in the sample, and *TN* is the number of true negatives. In general, for these metrics we seek a low false positive rate along with high precision and recall. For all models, we denote "positive" or fail by 1, and "negative" or pass by 0.

In addition to these numerical metrics, we also use the Receiver Operating Characteristic (ROC) curve as a visual basis of comparison. ROC curves are graphic illustrations of the performance of a binary classifier based on recall (TPR) and FPR. ROC curves can be built using probability-based classifiers by plotting TPR against FPR under different probability thresholds. For example, if a classifier assigns a probability of cancellation to a particular sample of 0.7, it will be labelled as a 1 for all thresholds greater than or equal to 0.7, and a 0 for all others. Thus, each threshold features different TPR and FPR values. An ideal ROC curve goes through TPR = 1 and FPR = 0, which means that under some threshold, we can achieve 100% accuracy in classification. A classifier based purely on a random guess is a straight line, while a practical classifier is somewhere in between these two cases. Figure 4-1 shows an example ROC curve. The area under an ROC curve (AUC) is also a useful overall metric for classifier performances with all possible thresholds. For an ideal classifier, AUC = 1; for random guesses, AUC = 0.5 [15].

Figure 4-1: Sample ROC Curve



4.2 Model Selection and Results

Based on initial research, we choose to evaluate two classifiers for this problem: random forest and Adaboost. This section describes each of the algorithms and some of their key features. It then describes the methods evaluated to treat data imbalances. Finally, this section shows the results of the model selection with accompanying criteria.

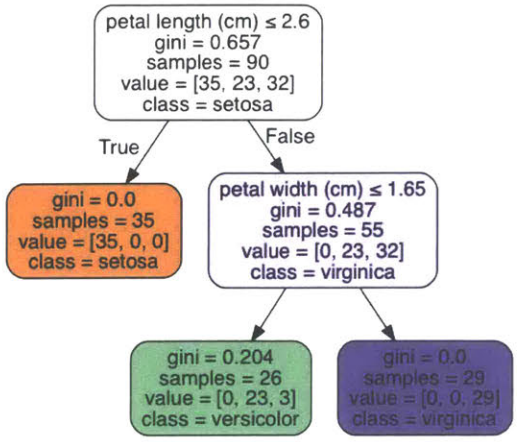
4.2.1 Description of Techniques

This section describes the algorithms used for predictive modelling. It discusses the basic underlying concepts of each as well as their general strengths and weaknesses. It also briefly discusses the parameters used to tune each model.

4.2.1.1 Random Forest

Random forest is one of the simplest and most pervasive machine learning ensemble methods. As Table 2-2 shows, it is a robust technique that is capable of working through numerous shortcomings in the available data to provide rapid and easily interpretable results. In the simplest terms, the algorithm works by building multiple decision trees and merging them together to build a more accurate and robust predictor. Figure 4-2 depicts a sample decision tree [16].

Figure 4-2: Sample Decision Tree



The decision tree is made up of multiple decision nodes that branch off to form a larger tree. At each node, the classifier chooses a feature from the data upon which to split and calculates the optimal value at which to perform this split to ensure minimal gini impurity in subsequent nodes. Gini impurity is a measure of how often a randomly selected sample from the data would be incorrectly labeled if it were randomly labeled according to the distribution of labels at the node. At all nodes, the classifier is seeking to minimize impurity, and if allowed, a decision tree classifier will continue to branch out into additional nodes until all nodes have a gini level of 0. Such a decision tree would be highly over-fitted to the data set and would likely be unreliable for use on any out-of-sample data. Random forest overcomes this danger by creating multiple unique decision trees and enforcing random selection of node criteria to ensure that each tree is not highly correlated to the others.

A random forest model can be tuned via a number of parameters. Many of these are related to how each tree is built, and these include, but are not limited to:

- Maximum tree depth: determines how many nodes deep, or how many decisions a tree is allowed to make
- Minimum samples per split: refers to the minimum number required to split an internal node
- Minimum samples per leaf: The minimum number of samples required at a terminal decision node.
- Minimum impurity decrease: A split will only occur at a node if the resulting gini impurity decrease exceeds a set threshold.

These parameters exist to help tune the accuracy of a model while preventing overly complex models that are over-fit to the training data set and incapable of handling the different value inherent in a test or out-of-sample set. A final parameter that is key to random forest is the number of estimators, or the number of trees in the 'forest.' This parameter is effectively a trade-off between run time and performance; as the number increases, model performance tends to increase while model run time also drastically increases.

4.2.1.2 Adaptive Boosting

Adaptive boosting is another ensemble method that builds off of base classifiers such as decision trees. In adaptive boosting, also known as "Adaboost," a base classifier is trained on the original data set. In subsequent base classifier training, instances misclassified by previous classifiers will be assigned larger weights. After training all the base classifiers, the ensemble classifier's final result is a weighted average of the base classifiers' output. The individual base classifiers may be very weak, but as long as the classification performances of each of them is slightly better than random guessing, the final ensemble classifier can be proven to converge to a strong classifier.

The parameters available in building an AdaBoost model include:

- Base classifier: decision tree, support vector classifier, etc.
- Number of Estimators: The maximum number of estimators at which boosting is terminated.
- Learning rate: there is a tradeoff between the number of estimators and the learning rate.

The main drawback to AdaBoost is that it is susceptible to noisy data and outliers, as it will attempt to address each outlier with added weight during each boost iteration. This drawback proves to be problematic with the Verizon datasets, as modelling results will show.

4.2.2 Treatment of Data Imbalances

In this study we must overcome the key challenge of an imbalanced data set. The minority class represents less than 10% of the overall dataset, with the majority class filling the remainder. Two data treatment methods are evaluated to alleviate the effects of this significant class imbalance. The first of these is random under sampling of the majority class. The technique involves randomly removing instances of the majority class from the training data until the desired class balance is achieved. Once the model is trained with the rebalanced data set, it can be tested on the imbalanced data set and evaluated for performance. The second technique is random minority oversampling, in which randomly selected instances of the minority class are duplicated until class balance is achieved. Each of these methods has potential shortcomings resulting from either willfully omitting data from, or introducing a high number of

duplicates into the training set. Adaptive Boosting, as mentioned previously, is an algorithm that is well suited for treatment of class imbalance and can be considered a third technique investigated to mitigate the impacts.

For the purposes of the following models, the minority oversampling and majority undersampling data treatments result in a balanced 50/50 split between the classes. Although other splits were contemplated and may in fact be more beneficial in certain use cases, the perfect balance is used in this thesis for simplicity.

4.3 Model Tuning and Results

This section outlines the results of predictive modelling to prevent waste within the Fios supply chain. It compares models built with the simple ordering dataset, as well as a more comprehensive second dataset. All of the predictive modelling work for this section was completed using Python 3 and the scikit-learn library.

4.3.1 Ordering Data Modelling

This section describes the actions taken to tune and evaluate the predictive model built using the ordering system data set. The random forest model is drastically improved through parameter tuning, but the result remains relatively poor in predictive performance by the selected performance criteria. AdaBoost tuning provides marginal benefit over the default model. Comparison of the tuned models shows that random forest outperforms AdaBoost for this data set at every FPR level.

4.3.1.1 Parameter Tuning

The first step in the modelling process is to tune the parameters. In previous sections, AUC, ROC curves, recall, and precision are discussed as effective criteria to evaluate model performance. However, only one criteria can be effectively employed at a time when tuning parameters. Since Verizon seeks low inventory holding costs, minimal inventory impact is a critical imperative for implementation of this model. Unfortunately, each false positive that results in unnecessary prevention of an item shipment leads to increased inventory requirements at local GWCs. Thus, it is critical to create a model that minimizes false positives while providing sufficient true positives to justify the effort. Therefore, tuning for this model is completed using a somewhat unique criterion: Partial AUC. As shown in figure 4-1, the x-axis for an ROC curve is the model's false positive rate. A partial AUC is derived from limiting the FPR to a certain maximum threshold and cutting the ROC curve off at that point. The partial AUC is then calculated for just that portion of the curve. Figure 4-3 depicts an example of a partial AUC. This criterion

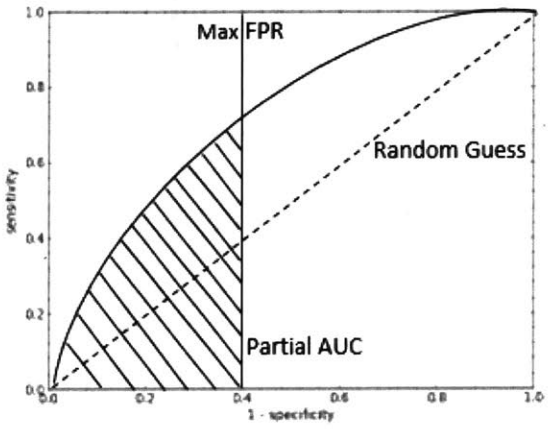
is used because it is only financially feasible to implement a model with a high detection rate at low FPRs. A model which, for example, approaches perfect predictability (TPR of 1) at an FPR of 0.5 but only gently slopes upward from the origin is not implementable in this business case. In order to allow proper comparison of partial AUC values across various maximum FPRs, we normalize the values based on the following equation:

Equation 4-2: Partial AUC Normalization [17]

$$Normalized\ Partial\ AUC = \frac{1}{2}\left[1 + \frac{Part.AUC - Min.\ Area}{Max.\ Area - Min.Area}\right]$$

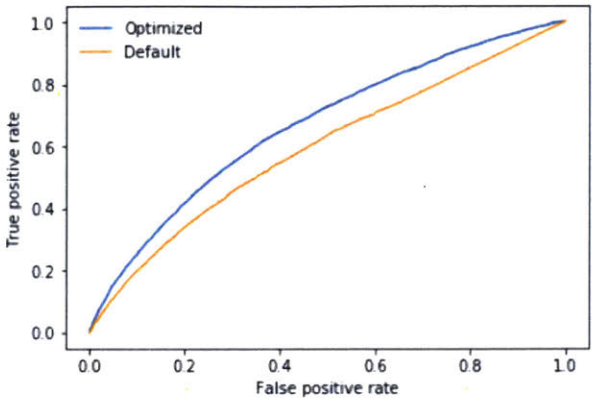
Where max area represents the maximum AUC value for the chosen FPR range. For example, for an FPR range from zero to 0.4, the max AUC is 0.4. The min area represents the minimum AUC value for the chosen range. For an FPR range from zero to 0.4, this value is 0.08, calculated by determining the area underneath the dashed "random guess" line for the portion of the curve in question.

Figure 4-3: Sample Partial AUC



In order to efficiently tune the model, we employ scikitlearn's built in gridsearchCV, which iterates model training and testing over every possibility of parameters within the input parameter grid. Multiple gridsearch iterations are required to narrow in on the optimal parameters. The data for this tuning is divided into training and test sets using scikitlearn's built in train_test_split function. The training set contains 70% of the data, while the test set comprises the remaining 30%. After employing these tools, the random forest model shows significant improvement with parameter tuning on the original data set. Figure 4-4 shows the ROC curve for the default random forest model and the tuned version.

Figure 4-4: Order Model Tuned Random Forest



The tuned parameters for this model are:

- Maximum depth: 20
- Minimum samples per leaf: 2
- Minimum samples per split: 10
- Number of estimators: 100

The AdaBoost tuning does not yield similar improvements, as tuning efforts are met with marginal improvements in partial AUC performance, as shown in Figure 4-5.

Figure 4-5: Order Model Tuned AdaBoost

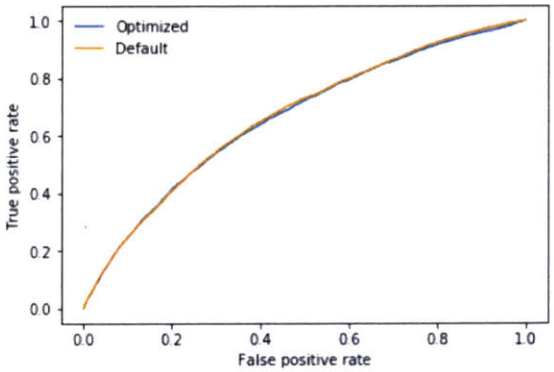
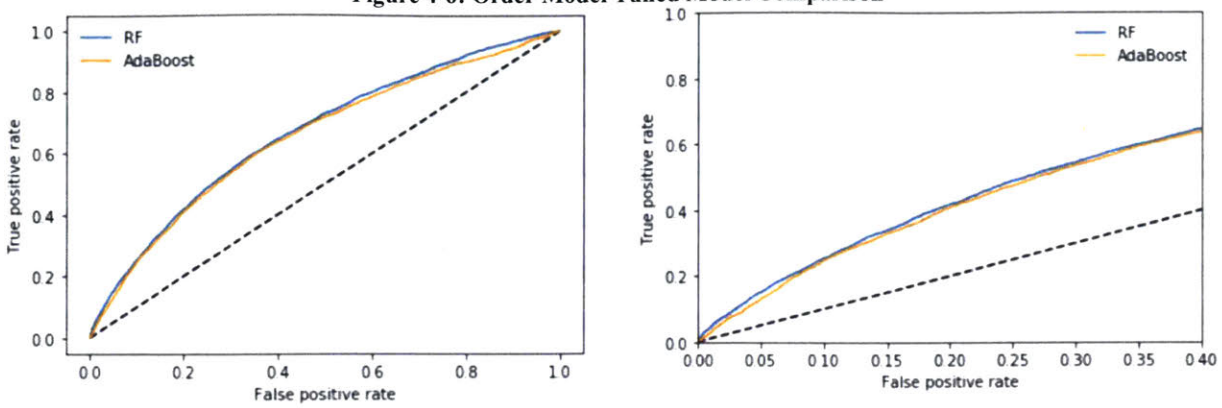


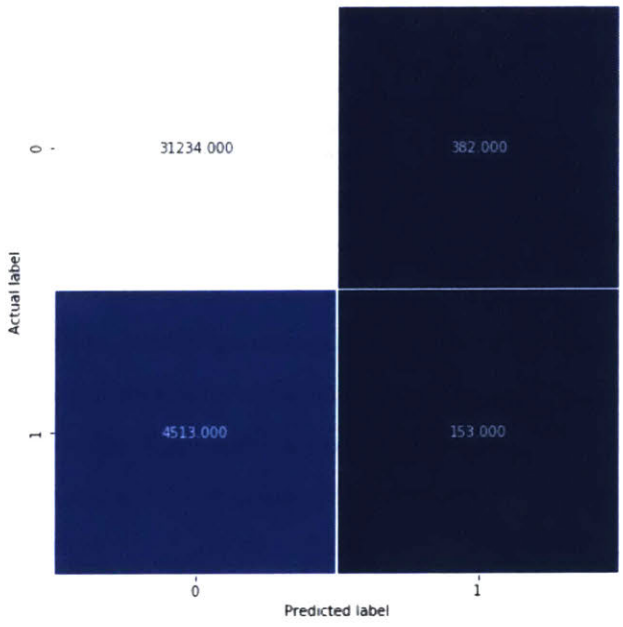
Figure 4-6 shows a comparison of the two tuned models on the base data set. The graph on the left depicts the entire ROC curve, while the right figure shows the ROC curve to the left of FPR=0.4. Although the random forest model performs slightly better within the partial AUC region and overall, the difference is marginal.

Figure 4-6: Order Model Tuned Model Comparison



Although these models built on the base data set show decent performance by AUC metrics, a confusion matrix at 50% confidence for the random forest model shows zero predictions of cancellations due to the skewed data set. To overcome this, it is necessary to experiment with the previously mentioned data treatments of oversampling and undersampling. Surprisingly, the AdaBoost model also exhibits minimal predictions. Figure 4-7 shows the confusion matrix for the AdaBoost model on the test data set. The result is heavily biased toward non-prediction in spite of the algorithm's general ability to handle such imbalance.

Figure 4-7: Order Model AdaBoost Confusion Matrix



4.3.1.2 Imbalance Treatment

Figure 4-8 shows the performance of the random forest model and the AdaBoost model on the base data set as well as the oversampled and undersampled data. The tuning program based on partial AUC for the random forest model is rerun for each new data set but the parameters remain the same. From these curves it is apparent that AdaBoost is not suited for artificially balanced datasets. In addition, it appears that random undersampling of the majority class has resulted in the highest performing model. Table 4-1 shows the key metrics for each of the distinct models at a prediction probability threshold of 0.5.

Figure 4-8: Order Model ROC and Partial ROC Curves with Treated Data Sets

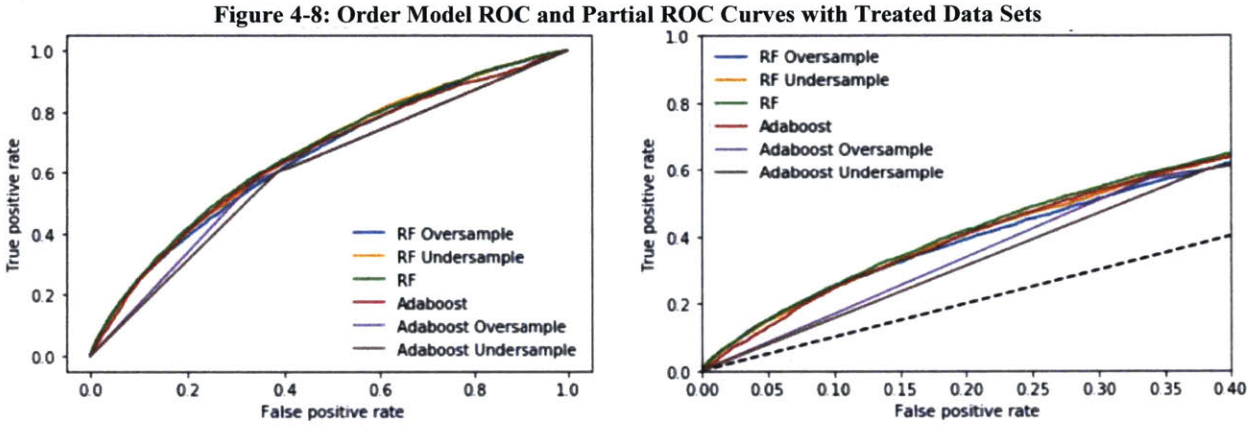


Table 4-1: Order Model Metrics with Threshold=0.5

Model Description	Partial AU	AUC	Sensitivity	False Positive Rate	Precision
Random Forest Raw Set	0.5498	0.6670	0.0002	0.0000	0.5000
AdaBoost Raw Set	0.5412	0.6533	0.0328	0.0121	0.2860
RF Oversampled	0.5448	0.6418	0.4835	0.2842	0.2007
RF Undersampled	0.5449	0.6579	0.6198	0.3840	0.1924
AdaBoost Oversampled	0.5409	0.6481	0.5705	0.3376	0.1996
AdaBoost Undersampled	0.5174	0.6335	0.6020	0.3854	0.1874

From this data, it is clear that none of these models are directly usable in our business scenario. The random forest model derived from the undersampled data exhibits the strongest performance in terms of high sensitivity and partial AUC, but it suffers from a fatally high false positive rate. The raw data random forest and AdaBoost models boast low false positive rates, but that is largely due to their failure to predict any cancellations, as evidenced by the catastrophically low sensitivity.

4.3.2 Sales Data Modelling

This section describes the actions taken to tune and evaluate the predictive model built using the more comprehensive sales data set 2. Parameter tuning results in improvement for both the random forest and AdaBoost models. Similar to the ordering system data set, the random forest model outperforms AdaBoost by most evaluation criteria. Comparison of data imbalance treatments shows that the random forest model using raw data performs very strongly at a confidence threshold of 0.5. Ultimately, the strongest model within this business case is the optimized random forest model built with the raw data set, at a confidence threshold of 0.7. This model yields an extremely low FPR below 1%, while detecting almost 40% of all order cancellations.

4.3.2.1 Parameter Tuning

Parameter tuning for the sales data set follows the same general path as the ordering data set. The random forest and AdaBoost models are first tuned on the initial imbalanced data set, and the confusion matrices are analyzed to determine the effectiveness of each model. Data imbalance treatments are then applied to allow for functional models. Similar to the ordering model, the random forest sales model shows significant improvement with parameter tuning. Figure 4-9 shows the ROC curve for the default random forest model and the tuned version based on partial ROC optimization. Figure 4-10 shows the tuned AdaBoost model against the default compared to the default setting. This plot shows a good example of a model with a high partial AUC but a lower overall AUC. In this case, the model exhibits high performance within the desired range of low FPR, which makes it the preferred model in this use case despite its lower overall AUC.

Figure 4-9: Sales Model Tuned Random Forest

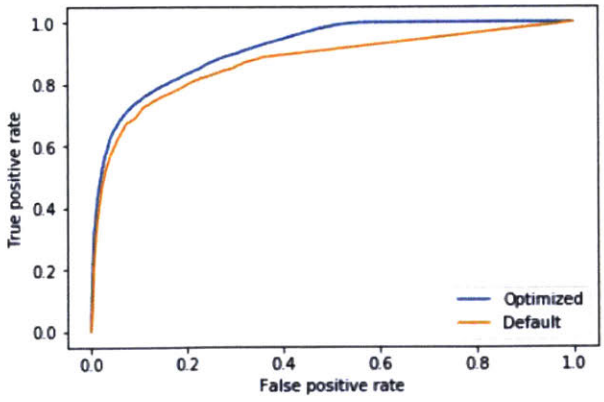


Figure 4-10: Sales Model Tuned AdaBoost

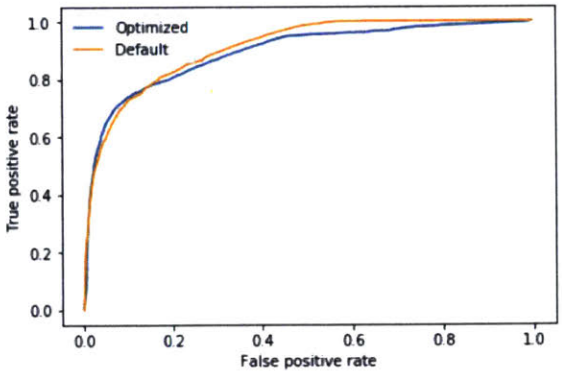


Figure 4-11: Sales Model Comparison of AdaBoost and RF

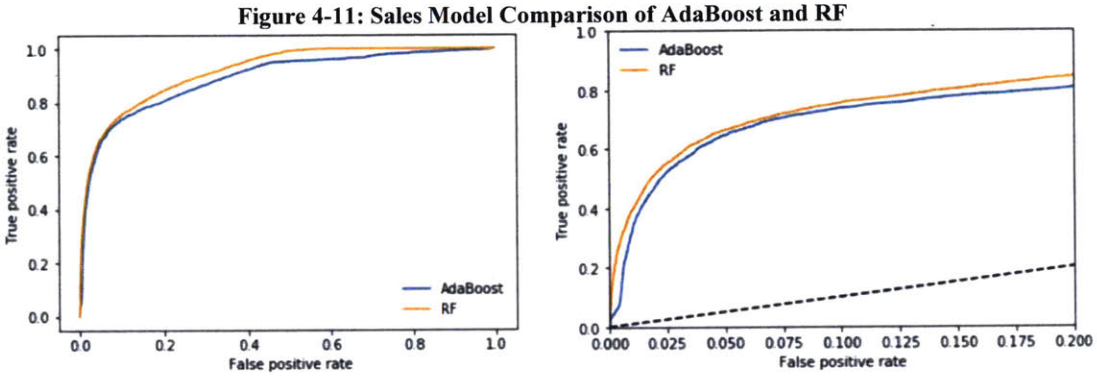
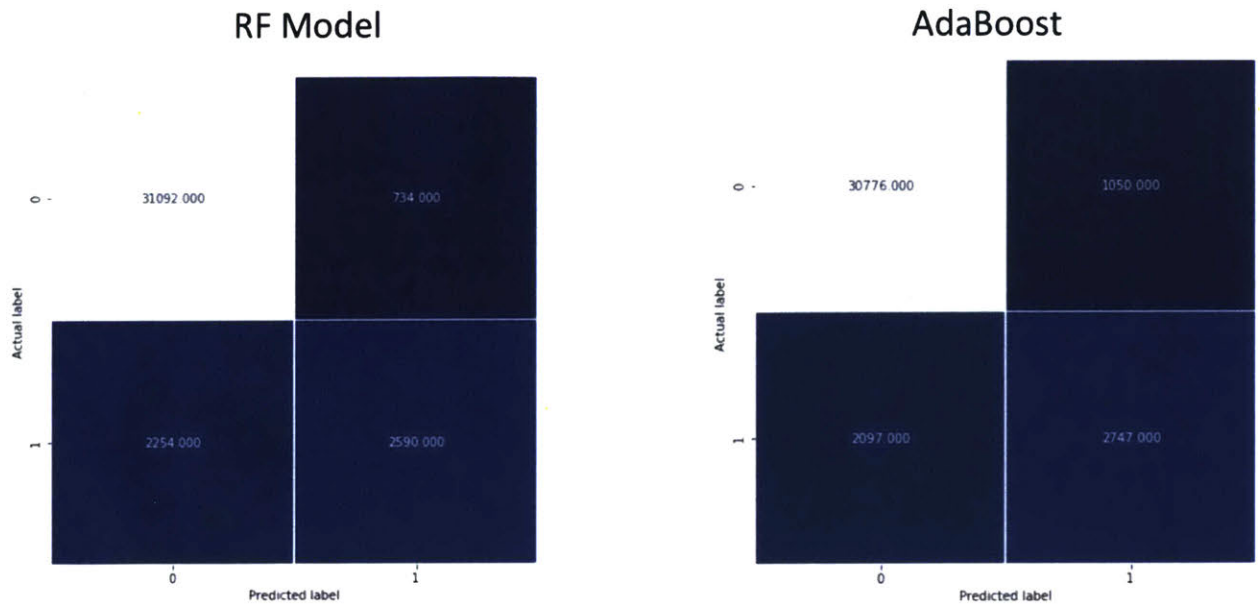


Figure 4-11 compares the random forest with the AdaBoost model. It is clear that, with the raw data set, the random forest model outperforms AdaBoost at every FPR value. In addition, unlike the sales data model, the random forest model built using the raw data shows excellent predictive capability. Figure 4-12 shows the confusion matrices for the random forest model alongside the AdaBoost model. Although the AdaBoost model predicts a higher percentage of the true positives in the test set, it also suffers from high false positive count. Both, however, perform well.

Figure 4-12: Sales Model Confusion Matrices



4.3.2.2 Imbalance Treatment

Despite the strong performance of the raw data models, it is still worthwhile to explore whether imbalance treated data sets can offer heightened predictive capability. Figure 4-13 shows the results for the tuned random forest and AdaBoost models built on oversampled minority and undersampled majority class data. Once again, this figure seems to indicate that AdaBoost does not perform well with artificially altered data. Overall, the raw data random forest model appears to perform slightly better over the length of the ROC curve, and the precise evaluation criteria are outlined in Table 4-2.

Figure 4-13: Sales Model ROC Curve Comparison

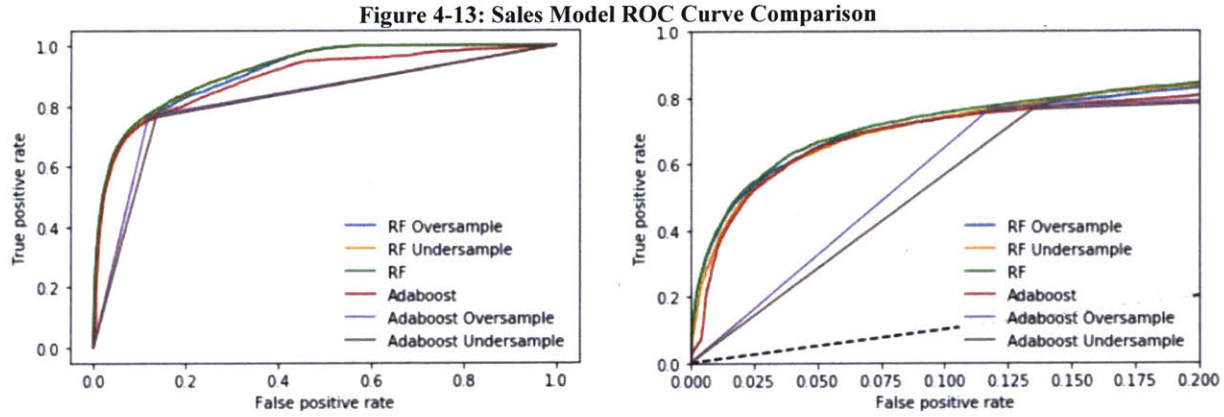


Table 4-2: Sales Model Metrics with Threshold=0.5

Model Description	Partial AUC	AUC	Sensitivity	False Positive Rate	Precision
Random Forest Raw Set	0.79487	0.91885	0.53468	0.02306	0.77918
AdaBoost Raw Set	0.77810	0.89110	0.56709	0.03299	0.72347
RF Oversampled	0.78826	0.91283	0.73906	0.10077	0.52748
RF Undersampled	0.78229	0.91553	0.80058	0.15192	0.44508
AdaBoost Oversampled	0.78227	0.90897	0.76734	0.11839	0.49659
AdaBoost Undersampled	0.75312	0.89993	0.76342	0.13473	0.46306

From this table it is clear that the random forest model displays excellent predictive capability combined with a low false positive rate. The imbalance-treated random forest models perform similarly in relation to AUC, but they suffer from higher false positive rates as a result of making far more cancellation predictions. The modelling results are even more compelling from a Verizon business perspective when the probability threshold is increased. Table 4-3 shows the evaluation criteria for each model at a threshold of 0.7. At this threshold, the base data set model successfully predicts almost 40% of cancellations with a false positive rate of less than 1%. The imbalance-treated random forest models are capable of detecting additional cancellations in the data set, but they each have higher FPR rates than the original. Notably, the AdaBoost models lack confidence in their cancellation predictions, and each model makes only a few cancellation classifications at this high confidence level, as seen in the extremely low sensitivity.

Table 4-3: Sales Model Metrics with Threshold=0.7

Model Description	Partial AUC	AUC	Sensitivity	False Positive Rate	Precision
Random Forest Raw Set	0.7949	0.9189	0.3708	0.0088	0.8647
AdaBoost Raw Set	0.7781	0.8911	0.0006	0.0000	0.7500
RF Oversampled	0.7883	0.9128	0.6117	0.0407	0.6957
RF Undersampled	0.7823	0.9155	0.6732	0.0601	0.6301
AdaBoost Oversampled	0.7823	0.9090	0.0002	0.0000	1.0000
AdaBoost Undersampled	0.7531	0.8999	0.0006	0.0000	0.7500

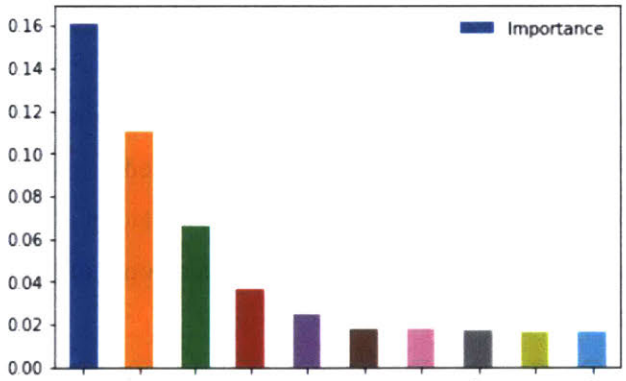
4.4 Application and Analysis

This section discusses the key takeaways from predictive modelling to detect order cancellations, covering both the ordering system model and the sales data model. The ordering system model is described as poor in its predictive capabilities yet insightful, as deeper analysis into the model's important features reveals a potential low-tech and easily implementable waste mitigation solution. The sales data model displays excellent predictive capability, but proves much more challenging to implement across the business.

4.4.1 Ordering System Model Application

The ordering system models exhibit relatively low performance, particularly in the low-FPR portion of the ROC curve. Each model displays a high FPR at the prediction threshold of 0.5, accompanied by low sensitivity and precision. There is no probability threshold that allows for any of the models to be directly used by the business in a prescriptive manner. The random forest models do, however, yield some significant insights primarily due to the model's feature importances. Feature importance is a calculated measure that describes the impact of each individual feature on the model's decision making. For a random forest, each decision tree develops unique feature importances based on the probability of the model reaching a feature's node and the resulting decrease in node impurity. The importances from each tree are then averaged over the forest to provide the overall model feature importance list. Figure 4-14 shows the top ten features by importance.

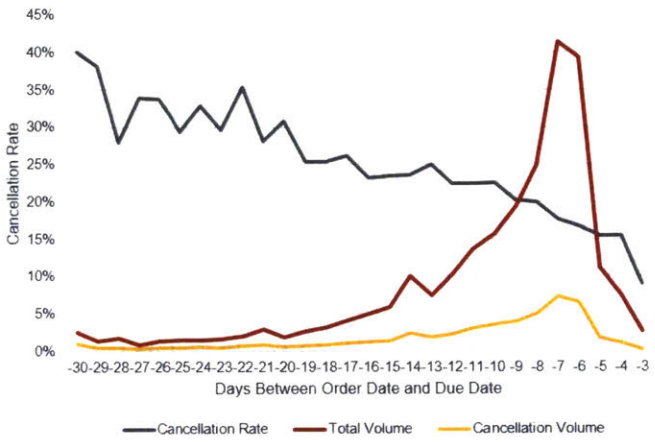
Figure 4-14: Order Model Top 10 Feature Importances



From the plot, it is clear that there are certain features that dominate the model. In fact, the model contains over 200 features after one hot encoding, but the top five features account for over 40% of the total importance. This finding by itself is not particularly surprising, and is in fact somewhat common. However, one particular feature from the top five list stands out and bears further discussion.

There are numerous avenues by which Verizon conducts sales, including third party companies, door to door salesmen, Verizon telephone agents, and web sites. Each of these unique avenues is identified in the data as a sales agency. There are over twenty such agencies in the data, and after one hot encoding, many of the smaller sales agencies become insignificant. The door to door sales agency, however, is the third most important feature in the model. This is particularly significant because door to door sales volumes are considerably lower than that of other channels, yet door to door is far more important in predicting cancellations. Further investigation into the matter reveals that door to door sales are, by themselves, less reliable sources for direct shipments. The overall cancellation rate for such sales is high and the probability of cancellation drastically increases as the number of days between the order placement and due date increases. Figure 4-15 shows the cancellation rate on door to door sales compared to the number of days between order placement and due date, as well as the affected volumes.

Figure 4-15: Door to Door Sales Cancellation Profile



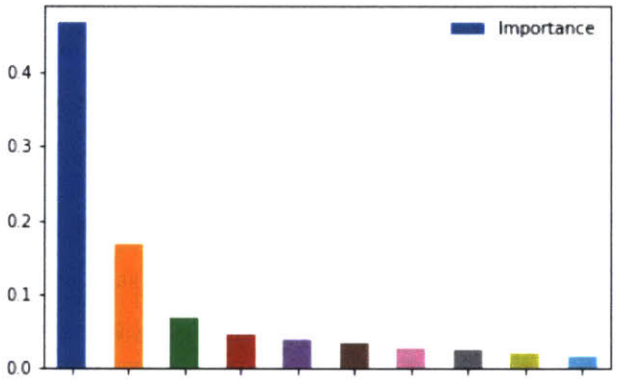
For additional context, the cancellation rate for sales channels other than door-to-door is less than 10%. In addition, inventory impact analysis indicates that the affected shipment volume is low enough to have marginal impact on GWC inventory levels. As a result, cancellation of all direct shipments on door to door sales is likely to save the Fios supply chain a significant amount of money by preventing wasteful shipments and stranded equipment.

Although the sales data models are not directly employable due to their high false positive rates, they remain useful by providing feature importance insight that allows for further targeted analysis.

4.4.2 Sales Model Application

Unlike the ordering model, the best sales model shows excellent performance characteristics. It detects cancellations at a high rate with an extremely low accompanying FPR. Unfortunately, compared to the ordering system data, the data used to build this model is not as simple to acquire in real time for rapid decision making. There are two potentially viable options moving forward to apply this model. The first is to analyze the sales model for important features and determine if these features can be made available in real time to the ordering data set to potentially drastically improve that model. Figure 4-16 illustrates that the sales data model is dominated by one particular feature, which relates to a customer's account tenure with Verizon. Given the simplicity of this feature, it is likely that there is a possible avenue to insert this information into the sales data to drastically improve the performance and viability of such models. For brand new customers, the value is simply zero, and for longer-tenure customers the data clearly exists within the Verizon system and it simply needs to be extracted in a timely manner. If such extraction is conducted, it is likely that the sales data models will become viable for prescriptive use within the supply chain.

Figure 4-16: Sales Model Top 10 Feature Importances



The second option for this data is to implement the model in a prescriptive manner using a more sophisticated IT approach. The challenge for such implementation is that, since the direct shipment decision is made during customer interaction, the model must exhibit the following attributes:

- Real-time extraction and joining of data
- Simple or automatic for sales personnel to use, and easy to disseminate
- Extremely short run time to display decision

During this project we partnered with a third party developer that specializes in real-time decision making using big data to develop a pilot program using the model as a baseline for prescriptive analytics. Investigations into the value and viability of this program are currently ongoing.

5 Conclusion

This section outlines the results of this project's waste reduction efforts within the Fios supply chain. Section 5.1 discusses generally applicable lessons regarding predictive modelling and process analysis. Section 5.2 outlines specific recommendations for Verizon based on the research in predictive modelling and process analysis. Section 5.3 provides recommendations for future research.

5.1 Generalized Lessons

This section highlights some of the key generalizable takeaways from the discussed research. It discusses lessons learned in both process analysis and predictive analytics.

5.1.1 Process Analysis Insights

This thesis investigates process mining as a technique to discover the reality of process flows within the supply chain. The technique proves to be extremely useful when the data allows. Analyses of flows within individual systems show clear dominant paths, and the additional insights provided such as resource usage and node duration can be extremely insightful for understanding the dynamics within the individual system. Problems arise, however, when disparate sources of data are integrated to form an overall view of the entire supply chain and its multitude of systems. Dissimilar formats and inconsistent quality across data sources render process mining difficult in this case. It may be possible, with significant effort, to normalize the data from the various sources to render a useful process flow, but that is a matter for future research.

Analysis of Verizon's RFID usage also yields a significant insight. Data integrity is crucially important in automated RFID systems. Although there are ways to establish fail-safes and self-corrections within the supply chain, automated RFID decisions are capable of exacting a significant toll if the data on which these decisions are based is unreliable. As mentioned in Sections 2-2, management decisions are only as good as the data on which they are based. Any business must ensure that their processes surrounding data transfer and storage are high integrity before endeavoring to incorporate significant RFID usage into their organization. This is particular important for business which operate in a 3PL enabled environment where third parties are frequently handling a business' equipment, and when necessary replacing damaged RFID tags. Communication with such 3PLs is extremely important to ensure the fidelity of the RFID tag data within the system.

5.1.2 Predictive Modelling Insights

This thesis introduces partial AUC as an optimization and evaluation criterion for predictive models. Partial AUC is extremely useful for creating a model with optimal performance in a specific portion of the ROC curve, and figure 4-10 illustrates a good example of this idea. The tuned model in this plot has a higher partial AUC, but a lower overall AUC than the alternative. Since this business case requires low false positive rates and a maximum FPR of 0.2, the overall AUC is ignored in favor of the partial AUC. Such a decision is likely to be applicable in any case where false positive predictions carry a significant penalty or cost.

The data imbalance treatments also yield an interesting insight. For both the ordering and sales data sets, the raw data set random forest models exhibited higher partial and total AUCs. This implies that the data treatment techniques applied for this thesis are sub-optimal. For the sales data set, the default prediction threshold 0.5 results in a confusion matrix with zero positive predictions due to the class imbalances. However, the indicators for classification are strong enough within the data that adjustment of the prediction threshold allows for use of the model in a predictive capacity. It is perhaps logical that a model built on randomly duplicated or omitted data may be less than ideal, and the idea of tuning a model's prediction threshold allows us to work around the class imbalance issue without such data manipulations. The AdaBoost model is also presented as a technique to overcome class imbalance within the data. Literature mentions that AdaBoost does not perform well on datasets with significant outliers, and this research confirms that idea. As mentioned previously, much of the data in this research contains outliers and inconsistencies. The effect of these issues can be seen in the reduced performance of AdaBoost relative the more robust Random Forest model.

5.2 Recommendations

A key goal of this thesis is to identify and mitigate sources of waste within the Fios supply chain. This section discusses specific waste reduction recommendations based on research in predictive modelling and process analysis.

5.2.1 Recommendations – Predictive Model

Although the predictive model built strictly using ordering system data shows poor performance by business- relevant evaluation criteria, the model's feature importance yields insight into opportunities to reduce waste. Primarily, further analysis of the door-to-door sales channel indicates that shipment of equipment on such sales is rarely a prudent course of action. Although the cancellation rate drops with the difference between order placement and order due date, even the lowest value is over double the overall

order cancellation rate from other channels. Based on this, there are two potential courses of action regarding door-to-door sales: Outright cessation of shipments on such orders, or cessation of shipments for door-to-door orders when the time between order placement and due date exceeds seven days. The first option prevents all of the waste associated with door-to-door shipments, but may have unforeseen inventory impacts. The rudimentary impact analysis neglected variables such as seasonality, regional preferences, and order variability, and there is potential that such a policy change might heavily impact specific GWCs, despite minimal overall impact on population averages. The second option is likely to prevent over 70% of the wasteful shipments associated with door-to-door sales without the elevated risk of inventory impacts. Either option is likely to result in significant waste reduction within the Fios supply chain. Implementation of either option should be relatively simple, as it could be disseminated to sales agents as a company policy change, effective immediately. It is recommended that either policy option be implemented as soon as possible in a limited pilot to test the impact and savings opportunity in practice.

There is also further opportunity to use the most effective model in a prescriptive manner to prevent undesirable shipment results. Although the ordering data alone does not provide sufficient predictive capability to merit further efforts, it is clear that augmentation of ordering data with key sales data provides excellent predictive performance. Efforts in this area should be focused on (1) adding key data such as account tenure into the ordering system data database or (2) real-time integration of the different data sources in a manner that allows a sales agent to provide a rapid decision on whether to direct ship or not. Implementation of such a prescriptive model across the disparate sales agencies may be difficult and it may be worthwhile to target specific troublesome order sources that have higher than average rates of cancellation.

5.2.2 Recommendations – Process Changes

Significant efforts have been made over the course of this project to understand the underlying processes within the Fios supply chain. This work reveals a complex underlying IT infrastructure that is a result of numerous mergers and acquisitions over the course of decades. One major problem within this network is data inconsistency across systems. Despite the fact that a single order requires input from six separate systems, it appears that there is very little data validation that occurs when information is transferred from one system to another. One specific recommendation on this matter is that instead of 'pushing' data from one system to the next without notification of receipt, each transaction should be validated. A validation process would likely result in slower transaction times or higher computing power requirements, but it would pay dividends in data consistency among systems.

An additional challenge within the supply chain is the use of RFID at GWCs. This process, outlined in Figure 3-6, is simple and logical but assumes perfectly-associated tags on all equipment to run smoothly. As this research has found, this is not always the case. As a result, physical GWC inventory counts can be misrepresented within the ERP system, leading to erroneous replenishment orders and some loss of awareness of the true state of the system. In the future, Verizon plans to increase the use of RFID across its supply chain and it is imperative that the identified tagging errors be rectified. Many of the problems with Verizon's RFID usage are related to data transfer challenges from 3PL suppliers. It is recommended that the process for such data transfer is refined to allow for increasingly accurate RFID tag data transmission and validation for each device.

5.3 Recommendations for Future Research

This research focused on analyzing Fios supply chain internal processes and on predictive analytics to prevent waste. Work on both lines of effort yields significant opportunity for future research.

5.3.1 Future Research – Predictive Model

Predictive modelling for this thesis is conducted with the general knowledge that any model implementation must limit false positives and the resulting inventory cost impacts. There is an opportunity to enhance the optimization of predictive models by incorporating real-world costs and benefits into parameter tuning. By incorporating model implementation costs, inventory impacts of false positives, true positive cost savings, and other related monetary impacts, it is possible to create an optimization model that analyzes predictions at each prediction threshold. This allows the user to simultaneously identify the optimal prediction threshold for each model and analyze the expected cost or benefit of its implementation. Such a model requires intimate knowledge of an organization's cost structures but may reap significant benefits in identifying ideal parameters and justifying the project to key stakeholders.

In addition, predictive efforts for this project focused entirely on direct shipments of equipment to customer premises on new technician install orders. This order type, although significant, only represents a small percentage of total shipments within the Fios supply chain. There is likely some potential to use predictive modelling for internal processes, such as GWC replenishment orders, that would result in cost savings across the supply chain.

Furthermore, the ability to implement prescriptive models in a customer-facing environment is somewhat limited by the challenge of real-time decision-making at the sales agent level. Further research

on methods to disseminate and implement prescriptive models under such circumstances is necessary to understand the challenges inherent in such an undertaking.

Finally, concept drift is briefly discussed in Section 2-2, but the idea is not addressed in this thesis. The existing models are built using batch data from a nine-month period and there is limited consideration of the temporal nature and potential time-related shifts in trends. Out of sample testing on more recent data batches shows good results, but it is likely that predictive capability will degrade over time. Further research into the potentially changing nature of Verizon's data and the resulting shift in prediction indicators is warranted. Incremental or online learning solutions may be investigated in future work.

5.3.2 Future Research – Process Analysis

The primary focus of process analysis in this thesis is on the use of process mining for root cause analysis. Due to the complexity and inconsistent interaction between the various constituent systems in the Fios supply chain, such efforts were generally in vain. It is relatively simple to conduct process discovery to understand the general flow of equipment and information, but we find it difficult to use such information in a systematic way to identify the sources of friction that may cause undesirable outcomes. Further research on process mining is required to identify the capabilities of the tool for statistical analysis and correlation identification.

In addition, analysis of the various stakeholders and incentive structures across the Fios network shows potential sources of conflict. One key example is highlighted in Chapter 3: the tendency of customer service representatives to ship a customer replacement equipment in order to keep the call duration short, despite the obvious waste of potentially shipping equipment unnecessarily. It may be worthwhile to conduct an analysis of the overall system dynamics to understand the various competing incentive structures within the business that potentially result in units striving independently for their 'local optima' instead of each part of the business collaborating in order to achieve the global optimum.

References

[1] "Verizon ends first-half 2018 with strong operating results," accessed February 15, 2019, from https://www.verizon.com/about/news/verizon-ends-first-half-2018-strong-operating-results.

[2] "Verizon Fios Coverage Map," accessed February 15, 2019, from https://broadbandnow.com/Verizon-Fios.

[3] Damodaran, Aswath. "The Little Book of Valuation: Characteristics of Mature Companies," accessed February 15, 2019 from http://people.stern.nyu.edu/adamodar/New_Home_Page/littlebook/maturecompanies.htm.

[4] Govindan, K., Soleimani, H., & Kannan, D. "Reverse logistics and closed-loop supply chain: A comprehensive review to explore the future." *European Journal of Operational Research*, Vol. 240, Issue 3, 1 February 2015, pp. 603-626.

[5] Toffel, M.W. "Strategic Management of Product Recovery," *California Management Review*, Vol. 46, No. 2, 2004, pp. 120–141.

[6] Gaur, J., Subramoniam, R., Govindan, K., Huisingh, D. "Closed-loop supply chain management: From conceptual to an action oriented framework on core acquisition," *Journal of Cleaner Production*, Vol. 167, 20 November 2017, pp. 1415-1424.

[7] Sahyouni, K., Savaskan, R.C., Daskin, M.S. "A Facility Location Model for Bidirectional Flows," *Transportation Science*, Vol. 41, Issue 4, November 2007, pp. 431-541.

[8] Deloitte. The hidden value in Reverse Logistics Point of view. Accessed February 17, 2019 from https://www2.deloitte.com/content/dam/Deloitte/be/Documents/process-and-operations/BE_POV_Supply-chain-strategy_20140109.pdf

[9] Hazen, B.T., Skipper, J.B., Boone, C.A. et al. *Annals of Operations Research* (2018) Vol. 270, pp. 201-211. Accessed from https://doi.org/10.1007/s10479-016-2226-0

[10] Fallon, Nicole. "Predictive or Prescriptive Analytics? Your Business Needs Both," Business News Daily, December 16, 2015. Accessed February 20, 2019 from https://www.businessnewsdaily.com/8655-predictive-vs-prescriptive-analytics.html.

[11] Deschmukhl, A., Kewlani, M., Ambegaokar, Y., Lanham, M.A. "Risky Business: Predicting Cancellations in Imbalanced Multi-Classification Settings." Accessed February 25, 2019 from https://mwdsi2018.exordo.com/files/papers/79/final_draft/MWDSI_Paper__final_submission.pdf

[12] Li, Suhong and Visich, John K., "Radio Frequency Identification: Supply Chain Impact and Implementation Challenges" (2006). *Management Department Journal Articles*. Paper 44. http://digitalcommons.bryant.edu/manjou/44.

[13] Statista. "Projected size of the global market for RFID tags from 2016 to 2020." Accessed February 20, 2019 from https://www.statista.com/statistics/299966/size-of-the-global-rfid-market/

[14] Aalst, Wil M. P. Van Der and Schahramn Dustdar. "Process Mining Put into Context." *IEEE Internet Computing* 16 (2012): 82-86.

[15] Chen, H. Boning, D. (2019) "Machine Learning Approaches for IC Manufacturing Yield Environment." In *Machine Learning in VLSA Computer Aided Design (*pp. 175-199). Springer.

[16] Deshpande, M. "Supervised Learning – Using Decision Trees to Classify Data." Published by Zenva Academy. Accessed February 26, 2019, from https://pythonmachinelearning.pro/supervised-learning-using-decision-trees-to-classify-data/

[17] McClish, D.K. "Analyzing a Portion of the ROC Curve." *Medical Decision Making* Vol. 9, no. 3 (August 1989): 190–95. https://doi.org/10.1177/0272989X8900900307.

[18] Attaran, Mohsen. (2007). "RFID: An enabler of supply chain management." *Supply Chain Management Journal*. 12.