A DATA-DRIVEN APPROACH TO REDUCE FOOD WASTE FOR A CONSUMER GOODS COMPANY

by
AFŞIN SANCAKTAROĞLU

Submitted to the Sabancı Graduate Business School
in partial fulfilment of
the requirements for the degree of Master of Science in Business Analytics

Sabancı University
June 2021

A DATA-DRIVEN APPROACH TO REDUCE FOOD WASTE FOR A CONSUMER GOODS COMPANY

Approved by:

Date of Approval: June 29, 2021

ABSTRACT

A DATA-DRIVEN APPROACH TO REDUCE FOOD WASTE FOR A CONSUMER GOODS COMPANY

AFŞIN SANCAKTAROĞLU

Business Analytics M.A. Thesis, June 2021

Thesis Advisor: Asst. Prof. Dr. Burak Gökgür

Thesis Co-Advisor: Assoc. Prof. Dr. Ayşe Kocabıyıkoğlu

Keywords: food waste, newsvendor, perishable inventory, machine learning, quantile regression

Today, the prevention of food waste has become a very significant issue for a sustainable future. In this study, an inventory planning process that will minimize both inventories and lost sales costs and indirectly food waste was studied by analyzing the sales data of a perishable product whose demand is random. The newsvendor problem has been adopted because it is a widely used perishable inventory management problem where the demand is uncertain. The traditional newsvendor problem is implemented on the assumption that the demand distribution is known. However, in reality the true demand distribution is unknown. Therefore, a data-driven and integrated solution method is used in our study by using machine learning models and quantile regression methods that do not require demand distribution knowledge. In the study where we use traditional demand forecasting methods and sequential demand estimation and optimization for comparison, we find that both the integrated demand estimation and optimization methods and machine learning methods perform better than their counterparts.

ÖZET

BİR TÜKETİCİ MALLARI ŞİRKETİ İÇİN GIDA İSRAFINI AZALTMAYA YÖNELİK VERİ ODAKLI BİR YAKLAŞIM

AFŞİN SANCAKTAROĞLU

İş Analitiği Yüksek Lisans Tezi, Haziran 2021

Tez Danışmanı: Yrd. Doç. Dr. Burak Gökgür

Tez Danışmanı: Doç. Dr. Ayşe Kocabıyıkoğlu

Anahtar Kelimeler: gıda israfı, gazete satıcısı, bozulabilir envanter, makine öğrenmesi, kantil regresyon

Gıda israfının önlenmesi sürdürülebilir bir gelecek için günümüzde çok önemli bir konu haline gelmiştir. Bu çalışmada, talebi rassal olan çabuk bozulabilir bir ürünün satış verilerini analiz ederek hem envanter ve kayıp satış maliyetlerini hem de dolaylı olarak gıda israfını en aza indirecek bir envanter planlama süreci üzerine çalışılmıştır. Gazete satıcısı problemi, talebin belirsiz olduğu ve ürünün bozulabilir olduğu durumlarda yaygın olarak kullanılan bir envanter yönetimi problemi olduğu için benimsenmiştir. Geleneksel gazete satıcısı problemi, talep dağılımının bilinmesi varsayımı üzerine uygulanır. Ancak gerçekte bu mümkün olmadığı için, çalışmamızda talep dağılımı bilgisi gerektirmeyen kuantil regresyon yöntemi ve makine öğrenmesi modelleri kullanılarak veri tabanlı entegre bir çözüm metodu kullanılmıştır. Geleneksel talep tahmini metodları ve ardışık talep tahmini ve optimizasyon yaklaşımlarının da karşılaştırma için kullanıldığı çalışmada hem entegre metodun hem de makine öğrenmesi metodlarının daha iyi performans gösterdiği tespit edilmiştir.

ACKNOWLEDGEMENTS

There are many people I would like to thank during this thesis process. I would like to thank to my advisors Ayşe Kocabıyıkoğlu and Burak Gökgür for their support and assistance in the research process, to Emre Tekinalp for their trust in me and support in data, to Ekin Öznar, Ayşe Doğan, Engin Kalkan, Ufuk Peker, Beliz Aldatmaz, Gizem Avcı and Yalçın Büle for their efforts in data sharing, and Gökhan Becerikli for his support on the clarification of the problem.

I would also like to thank my friends Sergen Tuğ Toraman and Sabri Karagönen for their ideas and information sharing.

Lastly, I would like to express my gratitude to my family for their unconditional supports so far.

June 2021

Afşin Sancaktaroğlu

I dedicated this study to Mustafa Kemal Atatürk and his friends who are founders of Turkish Republic.

June 2021
Afşin Sancaktaroğlu

TABLE OF CONTENTS

LIST OF TABLES .... **x**

LIST OF FIGURES .... **xi**

1. INTRODUCTION **1**
- 1.1. Motivation and Objective 1
- 1.2. Problem Definition and Summary of the Results 2
- 1.3. Contributions 2
- 1.4. Structure of the Thesis 3

2. LITERATURE REVIEW **4**
- 2.1. Food Waste Literature Review 4
- 2.2. Newsvendor Literature Review 8

3. EMPIRICAL SETTING AND DATA **13**
- 3.1. Empirical Context 13
- 3.2. Problem Description 14
- 3.3. Data Description 16
- 3.4. Descriptive Statistics 17

4. METHODOLOGY **20**
- 4.1. Sequential Demand Estimation and Optimization 20
  - 4.1.1. Demand Estimation 20
    - 4.1.1.1. Traditional Methods 21
    - 4.1.1.2. Machine Learning Methods 22
  - 4.1.2. Optimization of the Demand 24
    - 4.1.2.1. Model-based Optimization 24
    - 4.1.2.2. Data-driven Optimization 25
- 4.2. Integrated Demand Estimation and Optimization 25

5. FEATURE ENGINEERING AND FEATURE SELECTION **27**

5.1. Feature Engineering ... 27
5.2. Model Selection ... 30
5.2.1. Best Subset Selection ... 30
5.2.2. Selected Features ... 32
6. RESULTS ... **35**
6.1. Demand Forecasts and Performances ... 36
6.2. Optimization and Cost Performances ... 37
7. CONCLUSION ... **39**
BIBLIOGRAPHY ... **41**

LIST OF TABLES

Table 2.1. Food Waste Research Area 5
Table 2.2. Newsvendor Literature Comparison 9

Table 5.1. Features used in the machine learning methods 28
Table 5.2. Models 32

Table 6.1. Forecast Performances based on RMSE, MAPE and MAE 36

LIST OF FIGURES

Figure 3.1. Data 16
Figure 3.2. Weekly Demand in Tonnes 17
Figure 3.3. Histogram of Weekly Demand 18
Figure 3.4. Seasonal Demand Distributions 19

Figure 4.1. Illustration of ensemble learning 23

Figure 5.1. Best Subset Selection (James, Witten, Hastie & Tibshirani, 2013) 31
Figure 5.2. Subset selection using C_p and AIC 33
Figure 5.3. Subset selection using BIC and adjusted R^2 34

Figure 6.1. Cost Performance Analysis-Train Set 37
Figure 6.2. Cost Performance Analysis-Test Set 37

LIST OF ABBREVIATIONS

AIC: Akaike Information Criterion

BIC: Bayesian Information Criterion

DNN: Deep Neural Network

EFB: Exclusive Feature Bundling

ERM: Empirical Risk Minimization

FIFO: First in First out

FRP: Function Robust Profit

GBDT: Gradient Boosting Decision Tree

GOSS: Gradient-based One-Side Sampling

JEO: Joint Demand Estimation and Optimization

KO: Kernel-Weights Optimization

LGBM: Light Gradient Boosting Machine

LSTM: Long Short-Term Memory

MAE: Mean Average Error

MAPE: Mean Average Percentage Error

MFNV: Multi-Feature Newsvendor

ML: Machine Learning

MSE: Mean-Squared Error

MV: Mean and Variance

MVS: Mean, Variance and Semivariance

OM: Operations Management

POS: Point of Sale

QR: Quantile Regression

RMSE: Root Mean Squared Error

RNN : Recurrent Neural Network

S-Norm: Seasonal Normalization

SAA: Sample Average Approximation

SEO: Sequential Demand Estimation and Optimization

SKU: Stock Keeping Unit

SOBME: Second Order Belief Maximum Entropy

WMS: Weighted Mean Spread

1. INTRODUCTION

1.1 Motivation and Objective

Food waste has reached alarming levels globally and nationally (Akkas & Gaur, 2021). The seriousness of this issue becomes more evident when we considering that one out of every three plates in the world is wasted, three children die of hunger every one minute, and sixty percent of wasted food can be recovered (TISVA, 2020). Even more than this, food waste has a direct negative impact on the environment, economy, and society (TISVA, 2020). Just a few of these harms include an increase in greenhouse gases released during the cultivation, processing, packaging, and transportation of the food we eat; higher storage and production costs of firms; and a greater number of undernourished people.

When we consider the dimensions and effects of food waste, we must be aware of what we can do to prevent it. Operations management (OM), one of the fields through which we can decrease waste, can play a remarkable role in reducing the food waste problem (Akkas & Gaur, 2021). Food waste is observed throughout the supply chain due to reasons such as inability to fully harvest due to a lack of workers in the field, insufficient storage and maintenance in logistics processes, excess orders due to inadequate planning, and overemphasis on freshness at the retail level (FAO, 2020). These problems have been addressed via solutions such as perishable inventory management, smart packaging, solutions to ugly produce, measurement and analytics, consumer behavior, and supply chain efficiency (Akkas & Gaur, 2021). In this study, we focus on perishable inventory management and support the decision making of the planning phase with the use of data.

1.2 Problem Definition and Summary of the Results

In this thesis, we examined the data of Tadım, one of the well-known companies in the packaged nuts and dried fruits sector in Turkey, and conducted a data-driven study aiming to both reduce food waste that occurs in raw material planning stage and maintain their order fulfillment level for customer satisfaction. Due to the nature of their products, it is important to reach the consumer in a short time after the packaging process, as they are susceptible to deterioration. The company plans its production weekly and supplies raw materials accordingly. In an environment in which demand is uncertain, it has to work with safety stock to avoid losing its customers. In this respect, decisions of the planning manager directly affect the profitability and food waste of the company. To assist in these decisions, we implemented a newsvendor model, a model that businesses use to determine the order quantities of perishable inventories only when the order ordered once at the beginning of the period is valid for a certain period. In the classical newsvendor model, the parameters of a certain demand distribution are estimated and an optimization problem is solved at a cost minimization based on this distribution. However, in most real-world applications the demand distribution and parameters are unknown, making the distribution assumption problematic (Scarf, 1958). The increasing availability of big data today can help tackle this problem and enhance the accuracy of inventory models in real-world cases (Huber, Müller, Fleischmann & Stuckenschmidt, 2019). This leads us to the data-driven newsvendor problem. According to our analysis, proven by performance measurement results such as MAE, MAPE, and RMSE, machine learning methods such as ours give superior results to traditional methods. Moreover, a jointly considered demand forecasting and optimization QR approach can provide an average 6% less costly solution than sequential demand estimation and optimization S-Norm and SAA approaches. In the best machine learning method, QR performs 8% and %0.5 better than SAA and S-Norm, respectively.

1.3 Contributions

Our study seeks to contribute to the literature in the fields of measurement and analytics and inventory management in a food waste context utilizing the data-driven newsvendor approach.

In this structure, we aim to answer the following questions in the study: (Q1) How much more effective are machine learning approaches to traditional approaches? (Q2) Is integrated demand forecasting and optimization more effective than sequential demand forecasting and optimization?

1.4 Structure of the Thesis

Section 2 provides a review of the relevant literature by examining studies in the field of food waste and data-driven newsvendor. In Section 3, we convey more detailed information about the company, its problems, and its processes. Section 4 introduces the methodology, machine learning methods used in the study, the integrated demand forecasting and optimization approach used to solve the problem, and the sequential demand forecasting and optimization approach that we used for comparison purposes. In Section 5, we describe the model and features used as well as reasons for their selection. In the final section, we implement the model and share the results obtained. Finally, we test our approach on a dataset provided by the company.

2. LITERATURE REVIEW

This section presents a review of the two main literature streams related to our research: i) food waste and ii) newsvendor problem. We aim that the reviewed studies add a holistic perspective to our problem.

2.1 Food Waste Literature Review

Food waste has found itself at the forefront of academic and political fields in recent years Reynolds, Goucher, Quested, Bromley, Gillick, Wells, Evans, Koh, Kanyama, Katzeff & et al. (2019). In this section, we follow the framework proposed by Akkas & Gaur (2021) and review the work on food waste in operations management under papers on supply chain technology, business model innovation, behavioral operations, supply chain logistics, and incentives and coordination in the supply chain. The information we have obtained in these studies in the field of food waste contributes to our understanding of the before and after of our current problem, to see what we can do during the development phase to solve the problem, and to provide insight into the difficulties that may arise. In our thesis, we contribute to the determination of the optimal stock level for a perishable product. Thus, we contribute to the reduction of food waste. We list and specify the studies according to aforementioned categories. We can use Table 2.1 to observe the categories and sub-categories of all the above-mentioned studies and our contribution to literature in a more detailed way.

Dusoruth, Peterson & Schmitt (2018) analyze all stages of the entire supply chain network in the food field to minimize food loss by identifying where food is wasted. By applying the food scrap and diversion factors from the literature to the available Minnesota data, they present the situation at every stage from the farm to the household. They find that food wastage in households is the highest with 43%, and

Table 2.1 Food Waste Research Area

Author(s)	Publication Year	Food Waste Category	Subcategory
Lee	2012	Business Model Innovation	Solutions to Ugly Produce
Akkas, Gaur & Simchi-Levi	2018	Supply Chain Logistics & Incentives and Coordination in Supply Chain	Supply Chain Efficiency & Within Firm Incentives
Li, Yu & Wu	2016	Business Model Innovation	Markdown Platforms
Yang , Xiao & Kuo	2017	Business Model Innovation	Measurement and Analytics
Akkas & Sahoo	2020	Incentives and Coordination in Supply Chain	Within Firm Incentive Issues
Belavina, Girotra & Kabra	2017	Business Model Innovation & Behavioral Operations	Donation Matching Software & Consumer Behavior in Retail
Akkas & Honbon	2018	Supply Chain Technology & Supply Chain Logistics	Inventory Issuance & Supply Chain Efficiency
Dusoruth, Peterson & Schmitt	2018	Supply Chain Technology	Perishable Inventory Management
Buisman & Haijema, Bloemhof-Ruwaard	2019	Business Model Innovation	Markdown Platforms
Broekmeulen & Donselaar	2019	Supply Chain Technology	Smart Packaging
Belavina	2021	Business Model Innovation & Supply Chain Logistics	Measurement and Analytics & Supply Chain Efficiency
Sancaktaroğlu	2021	Supply Chain Technology & Business Model Innovation	Perishable Inventory Management

the place where waste is made the least is the fields with 5%.

Broekmeulen & Donselaar (2019) empirically examine applications in the field of food to reduce food waste, increase the freshness of perishable products and increase profitability in perishable products in supermarkets. They employ a regression model which shows that the potential increase in productivity is very high. They determine the increase in-store shelf life, unpacking reduced waste by 34.8% to 43.1%, and increased freshness by up to 17%.

Akkas & Honhon (2018) examine the effects of the distribution of products with a fixed shelf life on profits and food waste to reduce costs and waste by observing the perishability problem of packaged products with supply chain processes. By designing an infinite-horizon dynamic programming problem over a stochastic demand on which they conducted an analysis on a real dataset, they found that FIFO (first in, first out) can reduce waste only in situations that are often difficult to implement. Heuristic approaches yield results that are 11.5% more efficient in optimality gap, 14% more efficient in pantry life, and 6% more efficient in waste in terms of profit, waste, and freshness. The study also reveals that FIFO, contrary to popular belief, fails in shelf-life management with cellar life results.

Buisman, Haijema & Bloemhof-Ruwaard (2019) examine how and how much discount and dynamic shelf-life practices can reduce food waste for retailers. In the study using the dynamic shelf-life simulation optimization model on stochastic demand, they find that the effects of discount and dynamic shelf life on retail, separately and together, are beneficial in preventing food waste. They also determine that the most effective method is the application where the dynamic shelf life and discount are conducted together.

Li, Yu & Wu (2016) study when and how often a company that replaces stocks periodically should make clearance sales. They use two myopic heuristics to set the right strategy and achieve the goal in this area. While the first of these heuristic approaches include only the inventory information, the other also uses the one-period remaining lifetime information. As a result of the study, they determine that the second model is more successful compared to the first model and that they obtain a result very close to the optimum level.

Yang, Xiao & Kuo (2017) examine pricing strategy, shelf space arrangement, and replenishment policy through the supply chain to reduce waste in perishable foods. They propose optimization methods aiming to maximize profit over stochastic demand for both single-item food and multi-item food supply chain settings. They find that discount rate directly affects the decision in the supply chain and optimal discount rate could be found by solving the suggested equations.

Lee (2012) studies how a wasted product can be transformed into a salable product to ensure sustainability. With the model created, the operational optimization and licensing strategies are combined and they show that profit maximization can be achieved by determining what their relationship is. They demonstrate that the firm's optimum operating strategy depends heavily on the cost reduction dimension provided by the by-product synergy process innovation, exemplified, by the cost of disposal or the cost of raw materials.

Belavina (2021) studies the effect of grocery store density on food waste in homes and markets to reduce carbon emissions. She creates a two-stage perishable inventory model to optimize the total amount of waste generated in markets and homes by balancing market conditions. The results of the study reveal that the growth in store density significantly reduces greenhouse gas emissions.

Belavina, Girotra & Kabra (2017) compare two revenue models, per-order and subscription, in the field of e-tailing, taking into account financial and environmental performance, to show the environmental impact of food waste. They conduct the study by establishing a model that brings together a company known to be successful with online and offline delivery and customers with uncertain demand. With per-order, the weight of the products ordered increase due to transportation costs and therefore food wastage, and with subscription, the frequency of orders increase as the annual transportation fee is taken as a one-time fee, but the use of vehicles increases, increasing greenhouse gas emissions. In addition, they find that increased shipping costs in the per-order situation lowers customer adoption and negatively affect the sales of these durable products. Finally, as the order frequency increases in the annual subscription, transportation costs increase, but a price increase here also reduces the number of customers.

Akkas, Gaur & Simchi-Levi (2018) examine perishable products by addressing the issue of channels and multiple locations since current methods are insufficient to determine the causes of expiration dates and food waste cannot be avoided In the study, they use cross-sectional models and they reach the following result: The statistically significant determinants of the expiry date of the product are determined as case size, aging of the supply chain, sales incentives of the manufacturer, replenishment workload and minimum order rule.

Akkaş & Sahoo (2020) provide an empirical study of the impact of workforce incentives on food waste and profits. They define penalty points for expiry to capture the cost of waste and show that, in this setting, profits can increase up to 1.4%. They also find that incorporating a penalty cost that is 2.5 the workforce profits prevents 37.7% of expiration related food waste.

2.2 Newsvendor Literature Review

Newsvendor problem is a model used to determine the order quantities that maximize the expected profit according to the demand in cases where the demand for short-lived products is valid only once at the beginning of the period and for a certain period (Porteus, 2002). If the order is placed too little, the incoming demands will not be met and the sale will be lost, and if too many are given, there will be little or no opportunity to sell the remaining products the next day.

Various studies have been conducted on this newsvendor problem in the past. However, we review data-driven studies which is the focus of our study. Due to the increase in machine learning efforts and the capacity of computers to solve more complex problems in a shorter time, more accurate solutions have begun to be produced for the newsvendor problem (Huber et al., 2019). In these studies in the field of data-driven newsvendor, we observe different machine learning methods, different optimization methods, and different data types. All these different approaches will contribute to our study to determine the most appropriate approach. They also make it possible to increase elaboration by using the information they provide when we want to go one step further in solving the problem. In our thesis, we contribute to examining the effects of integrated demand forecasting and optimization and sequential demand forecasting and optimization approaches. In addition, we contribute to the literature by using a combination of machine learning methods and optimization approaches different from other studies by making use of the future research topics we observed in the studies we found. Below, we summarize these studies and examine the data-driven approaches to the newsvendor problem. We can comparatively observe a summary of the studies mentioned above in Table 2.2.

Levi, Perakis & Uichanco (2015) study the data-driven newsvendor model by investigating the sample average approximation (SAA) approach where the demand distribution is unknown. In the study, they use a completely new method called weighted mean spread (WMS), they detect a remarkably tighter bound compared to previous SAA studies (Kleywegt, Shapiro & Homem-De-Mello, 2002; Levi, Roundy & Shmoys, 2007).

Table 2.2 Newsvendor Literature Comparison

Author(s)	**Publication Year**	**Methodology**	**Demand Estimation and Optimization**	**Problem Type**
Levi, Perakis & Uichanco	2015	SAA	Separated	Data driven
Saghafian & Tomlin	2016	SOBME	Integrated	Data driven
Natarajan, Sim & Uichanco	2018	SAA, MV Joint, MVS joint	Separated	Distribution Free
Hu, Li & Mehrotra	2019	FRM	Integrated	Data driven
Huber et al.	2019	Quantile Regression	Integrated	Featurized data driven
Ban & Rudin	2019	ERM, KO	Integrated	Featurized data driven
Oroojlooyjadid, Snyder & Takac	2020	DNN	Integrated	Data driven
Siegel & Wagner	2020	Asymptotic adjustment	Separated	Theoretical
Papanastasiou	2020	Two-sided learning	Separated	Data driven
Seubert et al.	2020	ANN	Both	Data driven
Xu, Zheng & Jiang	2021	Robust Optimization	Integrated	Data driven
Sancaktaroğlu	2021	Quantile Regression	Integrated	Data driven

Saghafian & Tomlin (2016) examine how to approach the newsvendor problem when there is only partial distribution information available. In their application on numerical experiments with the maximum entropy-based technique called second-order belief maximum entropy (SOBME), they compare the performance of SOBME with the sample average approximation approach (SAA) and they find that SOBME gives faster and more reliable responses to latent changes than SAA in the unknown true distribution. They also point out that SOBME outperforms purely data-driven approaches in this environment.

Natarajan, Sim & Uichanco (2018) compare the sample mean approach (SAA), mean and variance information (MV) with mean, variance, and quasi-variance (MVS) information to find a solution to the multi-item newsvendor problem under conditions of asymmetry and uncertainty. They show that expected profit loss decreases when the true distribution is heavy-tailed. In addition, they find that the model with partitioned statistics gives better results compared to the model that includes covariance information alone.

Hu, Li & Mehrotra (2019) take another perspective on the data-driven newsvendor problem under the unknown demand function by coordinating pricing and inventory decisions. Convex, concave, and general utility functions are included in the maximin framework, which is the optimization method, and the features and solution models of the function robust model (FRP) are discussed. The study is completed with simulated data and a real dataset and as a result, according to the experiments, there is a risk-reward tradeoff and FRP ensures a framework for that.

Huber et al. (2019) analyze the data-based newsvendor problem on a single product under unknown demand by making comparisons at various methods and levels in demand forecasting, stock optimization, service levels, and sample sizes. In a study on real data using quantile regression (QR), they determined that integrated demand estimation and optimization performed more effectively than the sequential demand estimation and optimization approach.

Ban & Rudin (2019) suggest an integrated demand estimation and optimization approach rather than a two-step process to the newsvendor problem. In this study, they suggest two algorithms based on the empirical risk minimization (ERM) principle, with and without regularization, and kernel-weights optimization (KO). After the implementation of the algorithms, they indicate that both ERM and KO algorithms outperform their benchmark studies by 23% and 24% respectively. Moreover, they prove that integrated approaches could eliminate the boosted errors caused by separated demand estimation and optimization models.

Oroojlooyjadid, Snyder & Takáč (2020) use deep learning algorithms to solve the newsvendor problem in specific circumstances which are unknown probability distribution and multi-feature data. They use the Deep Neural Network (DNN) method to solve the newsvendor problem because other approaches are considered unsatisfactory when the historical data are scant and/or volatile. Revised loss function which takes into account shortages and overages of the inventory is tailored to the deep learning algorithm for multi-feature newsvendor (MFNV) problem. As a result, this combination eliminates the multiplied error of applying demand estimation and optimization separately. Finally, experiments of this study complete on a real-data, and the results demonstrate deep learning offers well-satisfying results with the highly volatile demand data.

Siegel & Wagner (2020) study another newsvendor problem under a parametric set with the assumption of unknown demand whose probabilistic distribution's form is known however parameters are unknown. They focus on eliminating a systematic expected estimation error by providing an asymptotic adjustment. They examine simulation studies on exponential, normal and log-normal demand distributions. They find that statistically significant estimation errors can be eliminated by asymptotic adjustment.

Papanastasiou (2020) investigates the consumer effect on the newsvendor problem where a product is newly produced. He uses a procedure called two-sided learning which consists of information from both customer and the firm and he designs his research on a monopolist firm that preplans to sell a new product whose value is unknown. He finds that the future demand is precisely affected by social media. Finally, the experiments highlighted that if the effects of two-sided learning could not be considered relevantly, the cost of the product would be remarkably higher.

Seubert, Stein, Taigel & Winkelmann (2020) study a classical newsvendor problem. They aim to optimize stock levels and decrease food waste. To solve this problem, they conduct this research with real-world bakery chain data from Germany and they use an artificial neural network. Seubert et al. (2020) follow the concept that solving newsvendor problem with sequential demand estimation and optimization which is called SEO in this study and operating both demand forecast and optimization at the same time (joint demand forecast and optimization - JEO) as Huber et al. (2019), Oroojlooyjadid et al. (2020) do in their work. After finding the best model features and completing the study they reach that both SEO and JEO remarkably added value which is around 30% cost saving to the planning phase of the bakery chain.

Xu, Zheng & Jiang (2021) bring another perspective to the data-driven newsvendor problem with unknown distribution. To reach robust solutions with the construction of a protection curve for approximating the true density curve, a distribution ambiguity set with the nonparametric characteristics of the true distribution is set up by using data input. As a method of implementation, robust optimization is embraced and two main benefits are attained in this study over the traditional studies. First, it is specified that a reliable approximation to the true density and small variability of the profits is yielded with the protection curve. Second, quick refreshment of the protection curve by data input works well even the data is small.

3. EMPIRICAL SETTING AND DATA

In this section, we give information about the company where the study was carried out and the available data. In addition, we describe the problem, and then we indicate the actions taken to gather data and descriptive statistics.

3.1 Empirical Context

The research is carried out with Tadım from the food industry. Tadım, Turkey's leading brand in packaged nuts and dried fruits sector, was established in 1971. Today, the company serves 18 countries on 4 continents with its factories in Turkey and Germany with the vision "Packaging nuts and providing the customers with the best quality products in their freshest form at the most reasonable prices.". As can be understood from the vision, freshness is a priority for the company.

In Turkey operations, raw materials are supplied from certain domestic and foreign producers. Among these raw materials, sunflower seeds and pumpkin seeds are sent to the Gebze facility for other production processes after being eliminated and calibrated at Kırıkkale facility. Similarly, pistachios are sent to the Gebze facility to be packaged after being cleaned, separated from the outer shell, and roasted at the Gaziantep facility. Apart from these raw materials, all other products such as hazelnuts, walnuts, prunes are stored directly in Gebze and all their processes are completed there. In Gebze facilities, brine, roasting, and packaging processes are carried out depending on the product. Today, it provides service in Turkey with 26 sunflower seeds, 79 nuts, 7 nuts bars, 17 HORECA, and 6 tin types, with a total of 135 SKUs. After the products are packaged, two different ways are followed to deliver them to the end consumer. The first of these is to reach the end consumer directly through e-commerce. The other is carried out by delivery to more than 100 distributors all over Turkey. These distributors, located in 7 geographical regions, namely Marmara, Black Sea, Aegean, Mediterranean, Central Anatolia, Southeastern Anatolia, and Eastern Anatolia, convey their demands to the company 1-6 times per week according to the region and sales volume. Products are prepared for these demands. Prepared products are loaded into vehicles under the routes created in line with the orders. Since the facility is very close to Istanbul, all Istanbul orders are delivered to the distributors' warehouses within 24 hours at the latest after the request is received. The demands of other distributors are also met within 48 hours at the latest. Distributors deliver those products to sales points such as markets, grocery stores, and gas stations with their vehicles. Since preserving the freshness of the products is an important priority, it sells only as much product as it needs to the point of sale. The frequency of delivery of products to points such as markets, grocery stores, and gas stations is also determined according to sales volumes and does not exceed one week.

As a result of interviews with company managers, we determined sunflower seeds as the subject of this study. The most important reason for making this choice is that sunflower seeds are the most sold product. Sunflower seeds are supplied from various regions of the country and are stored in Kırıkkale facilities under optimal conditions. Sunflower seeds, which are not roasted in Kırıkkale facilities, are delivered to the Gebze factory with weekly planning in line with the needs of the company. Delivered products are roasted and packaged and delivered to the end consumer. While the product can be stored for a long time under suitable conditions before roasting, it should be consumed in a short time after roasting. Otherwise, the product starts to get stale and causes a loss of value or even food waste. Since the demand is not known in advance, the firm works with a certain stock level to maintain its prestige and to keep the customer service level at the highest level.

3.2 Problem Description

As stated in the previous section, for Tadım it is required to bring the appropriate amount of product from Kırıkkale facilities and put it into the roasting process to provide the targeted service level and to prevent waste. Producing more than needed causes additional inventory costs, while underproduction causes loss of profit. In the case of excess production, a cost increase occurs due to the longer use of production machines and the loss of value of the product in the measures calculated by the company. On the other hand, under-production may lead to loss of profit as the consumer cannot be reached, and greater losses as it creates the possibility of the consumer to adopt another product in the market. The demand is uncertain in the company where the planning is done weekly. The fact that the stocking decision is made before demand realization and the financial consequences of overstocking and understocking leads us to the classic newsvendor problem (Porteus, 2002). It is a single period newsvendor problem, as the amount to be requested from Kırıkkale is determined before the demand is realized in each period, and there is no additional replenishment option for the product during the selling horizon.

In this case, the main purpose is to determine the order quantity according to

$$\min_{q \geq 0} EC(q) := E[C(q;d)] \tag{3.1}$$

that will minimize the total expected cost, consisting of cost of overage and underage, which amount to b and h respectively where q is the order quantity, d is the random demand.

$$C(q;d) := b(d-q)^+ + h(q-d)^+. \tag{3.2}$$

If the cumulative demand distribution function (CDF) F of demand is known , then the well-known optimal solution is to determine the order quantity by

$$q^* = inf\{y : F(y) \geq \frac{b}{b+h}\}. \tag{3.3}$$

In practice, in most real-world cases the distribution of demand is unknown, and it is harder to solve than a known distribution of demand. Thanks to historical data, explanatory features, and new machine learning methods, the optimal order quantity can be determined more consistently (Huber et al., 2019).

In the classical newsvendor model, firstly the demand distribution is estimated and then the appropriate order quantity is determined. Data-driven newsvendor, on the other hand, reaches the optimal result by basing the inventory level directly on historical demand and characteristics. By removing the stages, the error is prevented from folding and better results can be obtained. Therefore, we focus on data-driven newsvendor.

3.3 Data Description

The study covers the demand for sunflower seeds in Turkey operations. We implement the proposed approaches using the weekly demand data of Tadım. The data include a period of 167 weeks from January 2018 to March 2021.

To obtain the weekly tonnage data for sunflower seeds in an accurate way, three different datasets are used. These datasets include daily order data, price & weight lists, and product lists. Daily order data includes all requests from distributors to the company on a package basis. The amount that could not be sent although requested can be accessed from here. The demand determined in this way is not censored demand. On the other hand, price & weight lists provide information to weight changes made in SKUs. Finally, the product list contains up-to-date information about the packaging of SKUs. We can also observe the detailed information about the content of the datasets from Figure 3.1.

Figure 3.1 Data

We next explain how we transformed the data. First of all, all price and weight lists are brought together. By matching these lists with the information in the product list, we reach the total weight of the products in a package. The information we need here is the daily weight information of a package of each SKU. For this reason, since the list we have consisted of only the dates when the weight change was made, the days in between and the weight information of those days are added by using the date and lag operations in Python. Then, the obtained list is matched with the daily order data. Finally, after removing products other than sunflower seeds from the data, which are not covered by the study, we convert the daily data to a weekly form.

3.4 Descriptive Statistics

Following the steps in the data description, we have a total of 167-week uncensored demand data for sunflower seeds. Figure 3.2 provides weekly demand values from 2018 January to 2021 March.

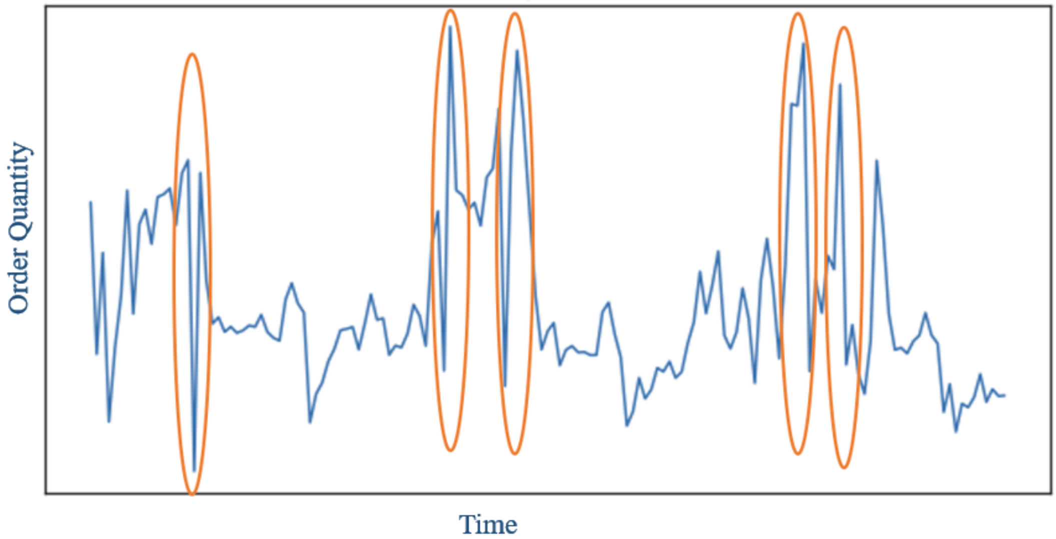

Figure 3.2 Weekly Demand in Tonnes

When we examine the Figure 3.2, we can say that the weekly demand is fluctuating. In addition, we can observe weeks of sudden decrease and increase in the areas between the ellipses in the figure. These periods are in the weeks before, during, and after the eid. Since there is special programming for the eid and there is no work during the eid, the demands are piled up before and after the eid. The production schedule for these special periods is also determined jointly with the sales and supply chain teams. On the other hand, it is also useful to see the distribution of demand after observing the transitions in demand between weeks. When we examine the demand distribution on the histogram (see Figure 3.3), we see that there is a right skewed distribution.

We applied the chi-square test to verify skewness of the data and the p-value of 7.47E-08 obtained as a result of the chi-square test. It shows the demand does not comply with the normal distribution. In addition, when we examine the data and in line with the information we obtained from the company, we see that there are differences in order levels between seasons. For this reason, we also examine the seasonal demand distributions which we present in Figure 3.4. We can infer that all seasons except autumn conform to the normal distribution, supported by the

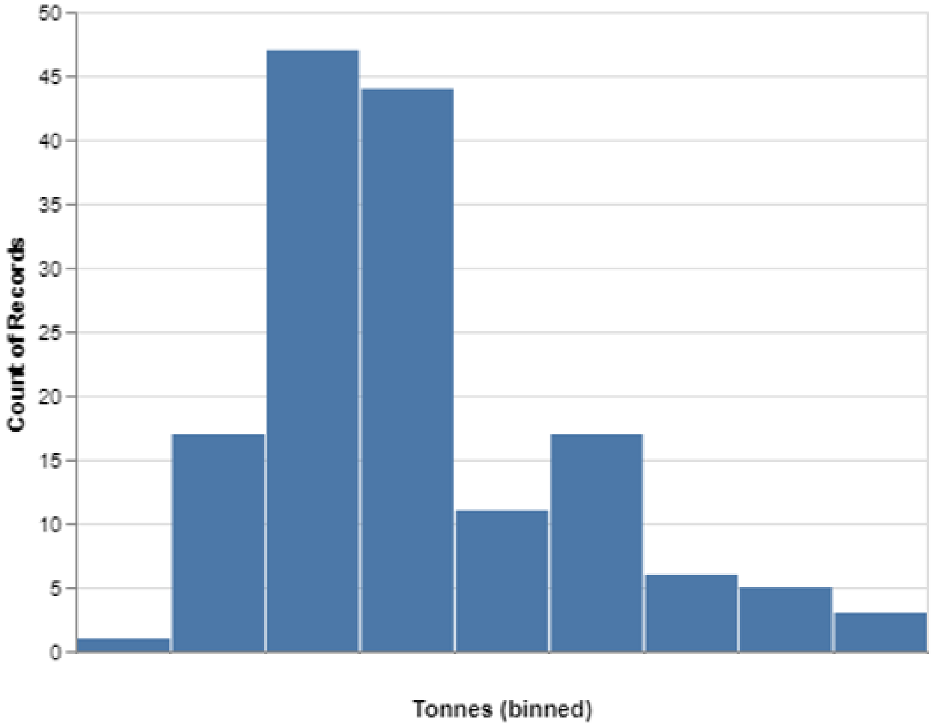

Figure 3.3 Histogram of Weekly Demand

results of the chi-square test. The p-values obtained in the chi-square test results for the spring, summer, autumn, and winter seasons are 0.176, 0.637, 0.001, and 0.549, respectively.

After a thorough examination of the dataset as we present in Figures 3.2,3.3, and 3.4, we excluded the outliers whose values above three standard deviations from the data. When we examine the weeks in which these outliers are extracted from the data, an important situation stands out. The reason for the high demand in these weeks coincides with the weeks in which the fulfillment rates of the orders from the distributors were below 70% in the previous week. As the distributors could not supply enough products in the previous week, they cause their demands in the relevant week to rise excessively and create anomalies. The main reason for the anomalies in both autumn and all seasons distributions stem from the orders that could not be met during the harvest period. This is the reason why these outliers were extracted from data.

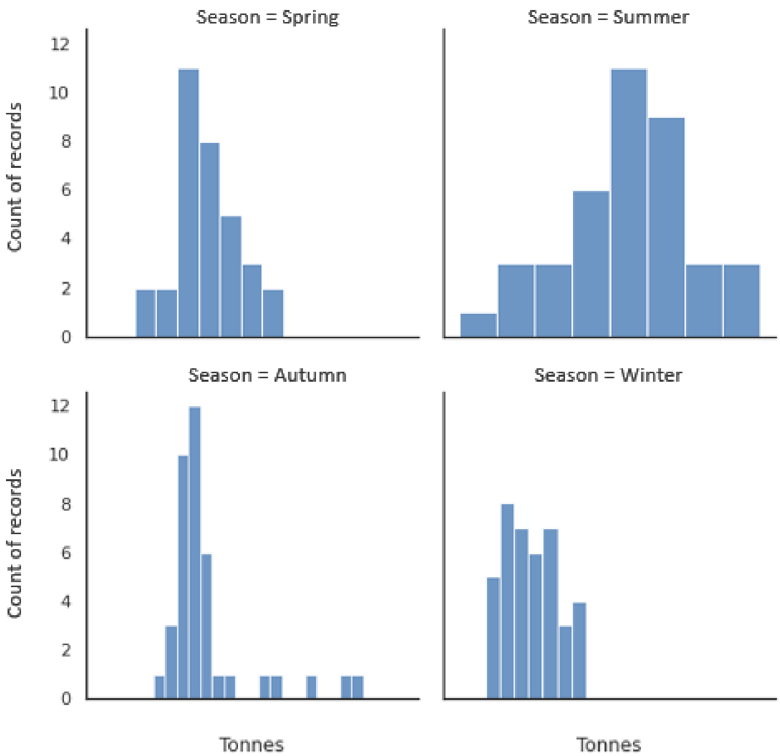

Figure 3.4 Seasonal Demand Distributions

4. METHODOLOGY

In this section, we present both benchmark and the machine learning methods that we will use in the analysis. We describe the sequential demand estimation and optimization starting with forecasting models. Then, we introduce model-based optimization which is based on demand distribution assumption and cost parameters (Silver & Pyke, 2017) and a data-driven optimization approach called sample average approximation (SAA). Finally, we describe the data-driven integrated demand forecasting and optimization models with the use of quantile regression (QR).

4.1 Sequential Demand Estimation and Optimization

This section provides a detailed description of how we construct the sequential approach, including two stages: i) estimating demand and ii) choosing the optimal order quantity. In this approach, we first estimate the demand for the focal product. Then, by making use of the demand structure, we determine the order quantity that minimizes the total expected costs.

4.1.1 Demand Estimation

Demand forecasting is an important element that guides the future planning of businesses such as investment and operating decisions. In general, underestimating the demand can lead to loss of customers, and overestimation can cause financial problems. In our case, the effect of overestimation may arise as a cost increase due to food waste and a liquidity problem for investment in other areas. Huber et al. (2019) underlined that forecasting models should have been taken into account when the structure of the demand data was uncertain. Today, with machine learning, better results can be obtained than traditional forecasting methods because of its ability to use high-dimensional data and detect nonlinear relationships.

In this section, we introduce both traditional methods and machine learning methods such as Linear regression, Random Forest, Xgboost, LightGBM, and Ensemble learning. The machine learning methods specified are chosen due to their different advantages, and these advantages are stated below in detail.

4.1.1.1 Traditional Methods

In this section, we present five different traditional approaches. The main reason for choosing these methods is that they are among the approaches used in the company.

1) Naive Method: The first of these, the naive method, means planning by foreseeing that the number of sales made in the previous sales period will be made. This means that the Lag1 variable in the data will be the expected result for this estimation method.

2) Seasonal Naive Method (S-naive): In the second method, the seasonal naive, the only difference from the naive method is that as its name suggests it takes into account the season. For example, while predicting the first winter week of the year, the demand is foreseen as much as the last winter order of the year.

3) Median: In the determined sample, the most common order quantity, that is the median, is predicted as future demand. It can be said that such an approach will be useful in cases where there is no fluctuation in demand which is not observed in our case.

4) Seasonal Median (S-median): Seasonal median, on the other hand, is the approach of grouping the available data according to the seasons and determining the most common order quantities as expected demand.

5) Moving average: Moving average is an approach that is made by taking the average of the determined number of orders before the demand to be predicted. In our case, the value chosen for this approach was determined as the value that gives the best result on the train set between 2 and 16 weeks.

4.1.1.2 Machine Learning Methods

In this section, we introduce six different machine learning approaches. We determine these methods by examining previous studies in the field of newsvendor problem.

1) Linear Regression: Linear regression is one of the most common supervised learning approaches. It has been applied for many years and it is a beneficial tool not only as a bouncing point for new approaches but also predicting a quantitative response (James et al., 2013). Linear regression indicates how the independent variables affect the dependent variable. The output of the linear regression can easily be interpreted.

2) Random forests: Random forest is a well-known tree-based method that is a combination of tree predictors. Each tree is built independently and their distribution is similar in the same forest (Breiman, 2001). This method has some similarities with boosting method; however, tuning and training are easier to apply (Hastie, Friedman & Tisbshirani, 2009). It can be applied to both classification and regression problems.

3) Xgboost: Xgboost is another tree-based learning method that uses a gradient boosting algorithm. The main advantages of Xgboost are its scalability and computational power (Chen & Guestrin, 2016). Like a random forest, it can also be applied to classification and regression problems. The main difference between these two is the combination of results and the tree-building method. As aforementioned, random forests' are built on independent trees while Xgboosts are built one tree at a time. Unlike random forests, results of Xgboost are combined throughout the process.

4) LightGBM: Light Gradient Boosting Machine (LightGBM) is a Gradient Boosting Decision Tree (GBDT) practice of the combination of Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) (Ke, Meng, Finley, Wang, Chen, Ma, Ye & Liu, 2017). GOSS eliminates some data points while keeping the prediction accuracy quite close to the original data which ensures time-saving (Ke et al., 2017). Likewise, EFB decreases computational load by making bundles to mutually exclusive features. These characteristics lead to more satisfactory results when the data is large, and a high feature dimension occurs.

5) LSTM: Long short-term memory (LSTM) network ensures finest outcomes with long time series data and disposes of the gradient problem (Kostadinov, 2018). Memory states are held in this sophisticated recurrent neural network (RNN) architecture

which helps to find relevant information, to understand what will happen in the next period of time.

6) Ensemble Learning: In ensemble learning, the goal is to combine simpler base models to benefit from their strengths and obtain a better prediction model (Hastie et al., 2009). Based on this, we can say that it provides an increase in performance by reducing the errors in the data. The error in machine learning can be expressed as

$$\epsilon = \sigma + \theta + t \tag{4.1}$$

where σ is variance, θ is bias and t is noise. Therefore, one or more of them should be improved to achieve high performance. Ensemble learning can be constructed by exploiting input features as Random Forest, exploiting train sets as Xgboost and LightGBM, and exploiting learning algorithms by LSTM which means some of the methods have already been introduced. In addition to these, it can be directly made by the combination of multiple models which can be observed in Figure 4.1.

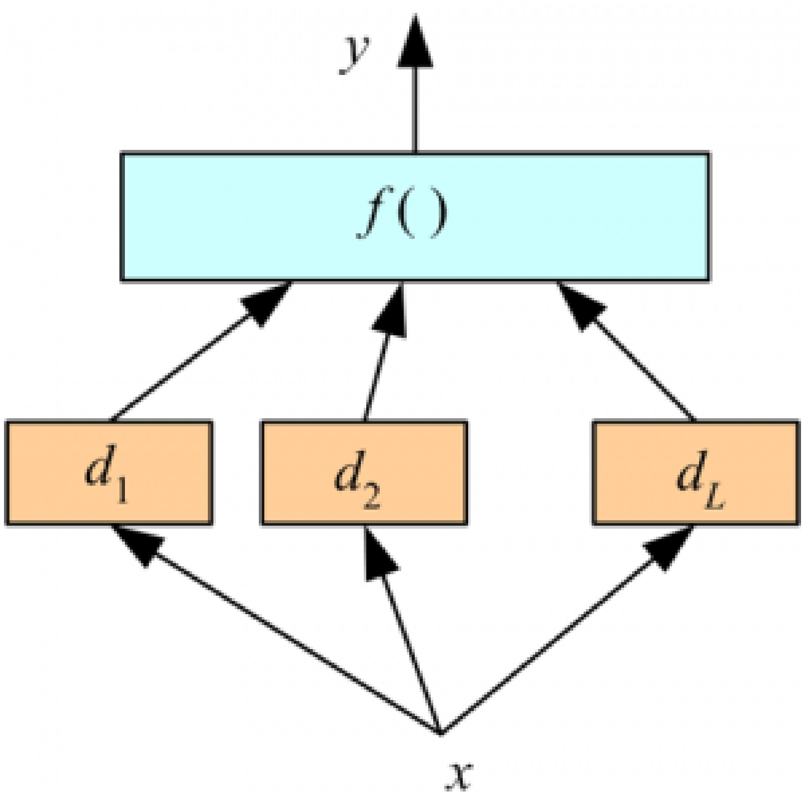

Figure 4.1 Illustration of ensemble learning

4.1.2 Optimization of the Demand

In the previous stage, we made the point estimation of the demand for the periods we determined with a margin of error. The second stage in sequential demand forecasting and optimization is inventory optimization to determine the optimal stock level after point estimation. After this stage, the amount of product to be produced is determined. In this section, we will present model-based optimization and data-driven optimization to better understand what is done after the demand forecasting.

4.1.2.1 Model-based Optimization

The approach, in which forecast error distribution parameters are determined using past prediction errors is called model-based optimization (Huber et al., 2019). Model-based optimization uses average and standard deviation information obtained from the demand estimation. Besides, overage and underage costs are also needed because the critical fractile is determined by these costs and is used to determine the optimal stock level. Although this calculation varies according to the type of distribution, in this study it was made by accepting the normal distribution . The reason for this is that when the demands are separated seasonally, the demand distributions correspond to the normal distribution for winter, spring, and summer seasons. The result of the ratio of $b \ / \ (b + h)$ (b underage, h overage costs) with respect to the normal distribution helps to find quantiles in the z table. By multiplying the value from the z table with the standard deviation, the amount to be produced is determined by adding it to the point estimation value. The assumption here is that the distribution estimate is correct. In many real-world studies, the true distribution has not been easily and consistently predicted (Yue, Chen & Wang, 2006). An error to be made here may have increased unbearably with the prediction error coming from the demand forecast in the first stage. For this reason, using data-driven approaches, which are more widely used recently, will give better results.

4.1.2.2 Data-driven Optimization

The sample average approximation (SAA) approach is one of the common and well-known data-driven decision-making approaches in the case of uncertainty (Bertsimas, Gupta & Kallus, 2017). The SAA method can be used in stochastic optimization problems, which are difficult to calculate although the distribution is known, such as two-stage discrete problems, and in problems such as newsvendor, which is easy to solve if the distribution is known, but in general, the distribution is unknown (Levi et al., 2015). In such a case, we can solve the problem as follows by making use of historical demand data and overage and underage costs regardless of the distribution

$$\min_{q\geq 0} \hat{R}(q;d(n)) = \frac{1}{n}\sum_{i=1}^{n}[b(D-q)^{+} + h(q-D)^{+}] \tag{4.2}$$

where D is the demand in i^{th} period and q is the forecasted quantity.

As it can be understood from the given equation, we can say that the quality of optimization depends on demand estimation performance and cost ratio. Therefore, forecasted quantity performance becomes a crucial issue for decision-making of the order quantity (Qi, Mak & Shen, 2020).

4.2 Integrated Demand Estimation and Optimization

Integrated demand estimation and optimization is a holistic method that the order quantity could be optimized straightly rather than separately estimating demand and optimizing order quantity levels (Huber et al., 2019). Since it is more difficult to define the demand model in high dimensions, it may result in the error of doubling in each step in two-step processes as separated demand estimation and optimization (Ban & Rudin, 2019). In inventory problems such as newsvendor, machine learning techniques can develop learning frameworks not only for prediction but also for integrated demand estimation and optimization (Qi et al., 2020). Formulation of this integration presented by Huber et al. (2019) as

$$\min_{\phi} \frac{1}{n}\sum_{i=1}^{n}[b(d_i - q_i(\phi, x_i))^{+} + h(q_i(\phi, x_i) - d_i)^{+}] \tag{4.3}$$

where $q_i(\phi, x_i)$ is the output of the ML method in period i with parameters and input variables x_i.

Quantile Regression (QR) is one of the methods to solve this problem that can be used when the assumptions of linear regression analysis are broken. In quantile regression, it is not necessary to know the demand distribution. This makes it a useful approach in the newsvendor problem setting. The QR enables the order quantity decision to be made as a result of fitting the chosen machine learning model according to the proportional status of the underage and overage cost $\frac{b}{b+h}$.

5. FEATURE ENGINEERING AND FEATURE SELECTION

In this section, we define new explanatory variables that are predicted to explain the variations observed in the data. Then, we choose the ones that will be included in the data to be used in the study according to the selection method we determined.

5.1 Feature Engineering

The addition of contextual information is seen as remarkable in the rapidly developing data-driven inventory management literature to eliminate unknowns (Qi et al., 2020). Therefore, we add new features to the dataset in light of insights received from the company and investigations made. You can observe these features and their data types from Table 5.1.

Lags are added to the dataset for using the demand information from previous weeks in machine learning methods and with that way we are be able to randomly split train and test data. The reason for going back up to 16 weeks is to catch the seasonal transitions.

Another variable, the holiday, is at the top of the periods that hold a remarkable place for the company. These holidays are religious holidays which are also explained in the table. Since shipments are stopped during the holiday period in the company, the number of working days decreases significantly during that week. The distributors are aware of this situation, they plan their orders accordingly before the holiday and increase their stock level. That is why it is important to keep this period, in which fluctuations were observed before and after, as three different binary variables which are *Holiday, after_holiday, and before_holiday.*

Ramadan is added as a binary variable. The reason for its addition is that the company manager has informed us that the sales of this period are lower than the

Table 5.1 Features used in the machine learning methods

Features	**Data Type**	**Description**
Lag1-16	Continuous	Previous weeks' demand in tonnes
Holiday	Nominal	Religious Holiday (Kurban or Ramazan) in that particular week (if it includes it is 1,otherwise 0)
Ramazan	Nominal	Ramazan in that particular week (if it includes it is 1,otherwise 0)
NewYear	Nominal	New year in that particular week (if it includes it is 1,otherwise 0)
Before_holiday	Nominal	1 week before the religious holiday (if it includes it is 1,otherwise 0)
After_holiday	Nominal	1 week after the religious holiday (if it includes it is 1,otherwise 0)
Bnewyear	Nominal	1 week before the new year (if it includes it is 1,otherwise 0)
Workdays	Ordinal	Number of workdays that company meets demand
LagStonnes	Continuous	Previous week's supply in tonnes
LagFullfilment	Continuous	Previous week's supply in percentage
LagChange	Continuous	The percentage change between last consecutive 2 weeks
LagChangeBinary	Nominal	The percentage change between last consecutive 2 weeks (if it is positive 1, otherwise 0)
Seasonality1-2	Continuous	Seasonal description

normal season. As a result of the observations, while sales in the first week of Ramadan are lower than in the previous week, there is an increasing trend starting from the second week.

The *new year* is another period in which sales are affected. Since the sales to the end consumer increased in the days just before the new year, the demands from the distributors are reflected in the previous week and the order quantities increase. In the week that includes the new year, sales are lower due to the shortage of working days and a large number of consumer purchases the products before the new year. Unlike the holidays, there is no unexpected increase observed in sales after the new year, so there was no need to add a separate variable after the new year.

The number of working days, which is among the reasons for the above variables, may vary due to company decisions or special days such as national holidays. This situation reveals that the number of working days is a factor that directly affects the future demand.

The other variable *LagStonnes* is the supply of the previous week, as indicated in the table. It was added as lag because it was not clear what was supplied without receiving the request. The main reason for its addition is the thought of closing the previous week's deficit in the demand to be realized in the next week when supply is low. This can happen especially in new crop periods. The distributor also makes more demands than usual, knowing that this will happen. Since we were not sure whether the lag supply or demand fulfillment rate would be more effective, another variable was added as the demand fulfillment rate.

Lagchange and *lagchangebinary* variables are added considering the direction and/or degree of sales volume change that could provide information about future orders.

Finally, there is the seasonality issue. The general trend from the new year to the end of summer is that if we remove the periods mentioned above, sales increase. For this reason, *seasonality* was added as a variable.

All these variables are added to make the changes in demand more explainable. Of course, some of the variables mentioned will be observed as no significant impact for reasons such as multicollinearity and will not be included in the model as their relative contribution will be seen as lower than the other variables. In the model selection section, we determine the optimal model that covers all variations in the data with adequate amount of variables.

5.2 Model Selection

Model selection is repetitive and exploratory because the range of model selection is often infinite and it is often impossible for analysts to know a possible combination that can provide sufficient accuracy and insight (Kumar, Naughton & Patel, 2015). In addition, model selection is made to predict the performance of several models to reach the best possible alternative (Hastie et al., 2009). These definitions imply that there is a reason for using a certain number of variables instead of using all the variables in the data. Model complexity decreases the bias of the train data, but the variance increases (Hastie et al., 2009). When the same model is applied to the test data, the test error increases. This is called overfitting, which means the model has poor generalizability. Additionally, the fewer explanatory variables, the easier it is to interpret the model (James et al., 2013).

In general, the model selection process proceeds as follows. First, the most predictive set of predictions is created. Then, several algorithms are selected from a set of models, and finally, algorithm hyperparameters are adjusted for performance optimization. Since we will proceed through the best subset selection in our study, we will examine each of steps now.

5.2.1 Best Subset Selection

For each possible combination of explanatory variables, separate least squares regression is fit when you implement best subset selection (James et al., 2013). We can also observe the stages in detail in Figure 5.1.

While it is guaranteed to find the best model in this approach, its main drawback is the computational limitations. As the number of explanatory variables in the data increases, the number of models tested in the algorithm increases exponentially, which is exactly $2^p - 2^{p-1}$ times. Therefore, forward, backward, or hybrid stepwise selection methods are used instead of the best subset, nevertheless, none of these approaches can guarantee the best model.

When the application of the best subset method is completed, as stated above, it offers as many model alternatives as to the number of variables, however, it does not answer the question of how many variables should be selected. Various methods are

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots p$:

   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Figure 5.1 Best Subset Selection (James et al., 2013)

used to determine the optimal model. As aforementioned, train data results are not suitable for determining this. For this reason, the smallest residual sum of squares (*RSS*) and highest R2 results in train data are not enough for the determination of the optimal model. There are two widespread approaches which are the direct estimation of test error and indirect estimation of test error. Now, we will discuss both approaches by investigating C_p, Akaike information criterion (*AIC*), Bayesian information criterion (*BIC*), *adjusted* R^2, and cross-validation.

First, when we look at the indirect test error estimates, it can be seen that the main purpose is to eliminate irrelevant variables by adding factors that make the selection of complex models difficult. According to C_p's equation to predict the test MSE which normally equals *RSS/n* in the test set,

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2). \tag{5.1}$$

where d is the number of predictors in the model, and n is the number of observations in the training set. Additive $2d\hat{\sigma}^2$ is a penalty and it is added for underestimating the test error. The small value of test error is better that is why the model with the lowest C_p value should be chosen. Likewise, the *AIC* criterion is another estimator of prediction error and it deals with the trade-off between complexity and simplicity of the model. *AIC* is proportional to C_p which can be observed from the equation

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2). \tag{5.2}$$

Table 5.2 Models

Model	Inputs
7 -variables	Lag1, Lag2, Lag6, Lag11, Holiday, Aholiday, Bholiday
10-variables	Lag1, Lag2, Lag6, Lag11, Holiday, Aholiday, Bholiday, Lag-Fullfillment, NewYear, Bnewyear

Similarly, the lower the *AIC* criterion score the better model we have. Next, *adds* a logarithmic penalty to the *RSS* which is a Bayesian perspective to the indirect selection approaches given by

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + log(n)d\hat{\sigma}^2). \tag{5.3}$$

Final indirect estimation of the test error approach, the *adjusted* R^2, a popular approach where the model with the higher score is chosen, unlike the others. The *adjusted* R^2 is calculated as

$$adjustedR^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \tag{5.4}$$

where *TSS* is total sum of squares.

5.2.2 Selected Features

The best-subset selection method is used to determine the most appropriate variables in the model. However, when using the best-subset selection, $2^p - 2^{p-1}$ times more models are tested for each new variable added, and this may cause problems in computational limits. In order to prevent this, first, we run all variables together, then we eliminate the variables with a p-value above 0.05. The best-subset selection was made after the elimination of these variables with the p-value above 0.05. This situation eliminated over 536 million possible models, saving a great deal of time. In order to prevent overfitting, two models with 7 variables and 10 variables were selected as a result of both indirect approaches and the validation set approach on train and test data. You can observe these models from Table 5.2.

When we look at the variables, we can say that the recent demand and the holiday periods are remarkable in explaining the variation. In addition, it would not be wrong to interpret *LagFullfillment* and Lag1 as meeting *LagStonnes*' relative influence. Similarly, the inclusion of Lag1 and Lag2 together in the model may indicate that the variables *LagChange* and *LagChangeBinary* do not provide sufficient explanation.

On the other hand, if we look at the reason why the number of models is two, it is due to the different results of C_p, *AIC*, *BIC*, and *adjusted* R^2. Only *BIC*'s result suggests using a less variable model because it adds more penalty on models with many variables (James et al., 2013). As a result of the study, it was not clear which would provide the greatest benefit, so both models were implemented. You can also observe C_p, *AIC*, *BIC*, and *adjusted* R^2 results on tables. As can be seen from Figure 5.2 and Figure 5.3, while C_p and *AIC* suggest a 10-variable model, *BIC* suggests a 7-variable model. Considering the *adjusted* R^2 graph, since the values between 10 and 12 variables are very close to each other, 10 is preferred instead of 12.

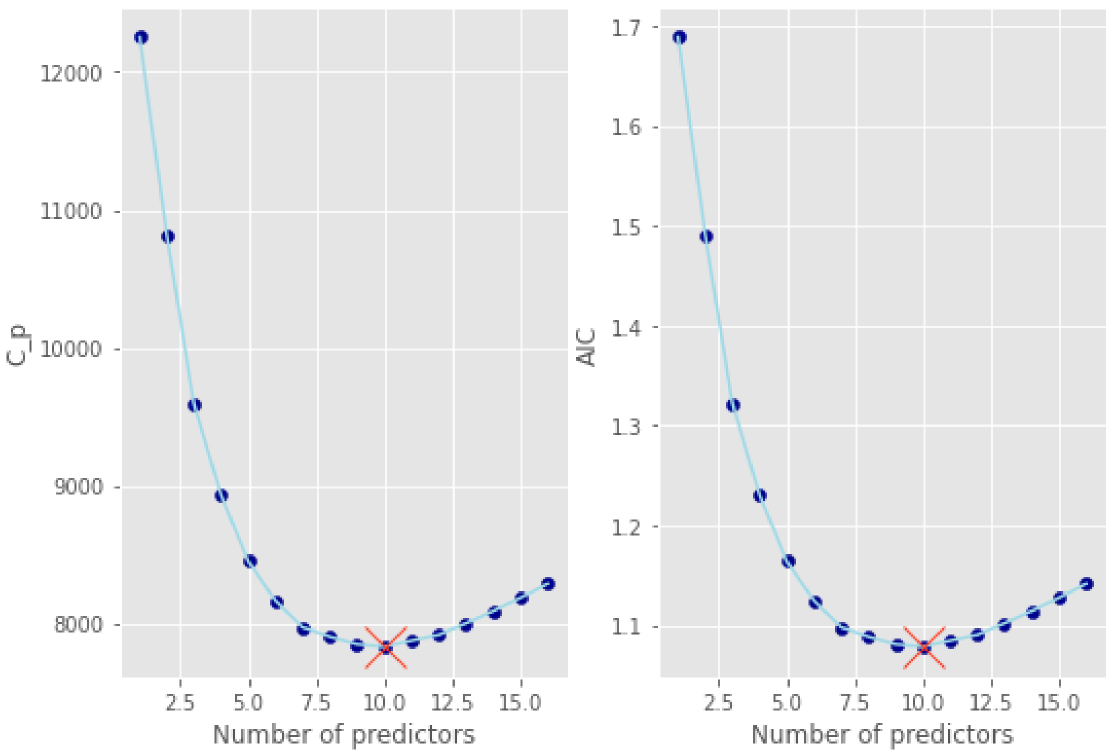

Figure 5.2 Subset selection using C_p and AIC

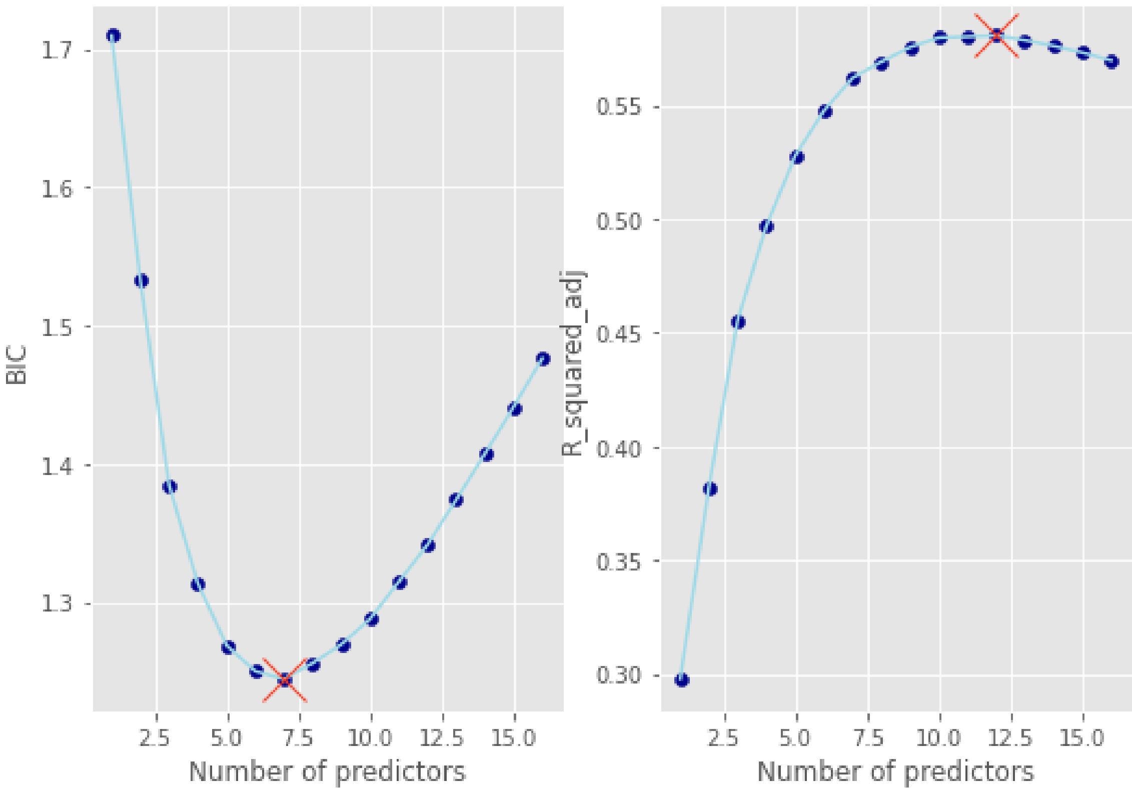

Figure 5.3 Subset selection using BIC and adjusted R^2

6. RESULTS

The results part is completed in two stages: evaluation of demand estimation performances and evaluation of objective function performances. We also compare the sequential and integrated demand estimation and optimization in this section.

To see the contributions of machine learning methods used in the study, we also use common traditional forecasting methods for comparison. After that, we observe the performances of all methods on the train and test set with the metrics we have determined. Two different models the one with 7 variables and the other one with 10 variables are conducted for linear regression, Xgboost, LgbmDT, and random forest using inputs determined as a result of the best subset selection method. Similarly, in these methods, the train and test set are determined randomly and the same observations were used in all of them. The data was divided into two as 70% train – 30% test set due to the low number of observations. Apart from this, to avoid overfitting, the number of leaves and learning rates in tree-based ensemble methods are determined with cross-validation. On the other hand, in LSTM, since the observations should be time series, train and test sets were divided into 70 and 30 according to the historical situation. Moreover, only tonnage data was used as input. In the ensemble method, which is made by combining the existing models, the method that gives the best result was chosen. We try to find the best ensemble model by trying different combinations of machine learning methods. As a result of these trials, the results obtained from LgbmDT and random forest with 7 variables are shown as the best results are achieved.

We conduct model-based optimization (seasonal normal distribution) and data-based optimization (SAA). For these approaches, we use overage and underage costs information which we get from the company and our demand forecast results. In addition, we will perform integrated demand forecasting and optimization with the QR approach. We then bring together all the results obtained and examine the cost performance of the objective function.

6.1 Demand Forecasts and Performances

To understand the quality of the forecast, their performance should be evaluated. There are various methods to do this. The reason for the variety is that different advantages will be explained below. We use these root mean squared error (RMSE), mean average percentage error (MAPE), and mean average error (MAE) in this study. Simply put, RMSE is the standard deviation of estimation errors. Since the magnitude of these errors varies from data to data, they only allow the comparison of those on the same scale. This shows that it is not correct to compare LSTM and other methods over RMSE in the study. MAE is also scale-based, just like RMSE, but here is a measurement made on the average of errors. On the other hand, MAPE is the scaled version of MAE and allows the comparison of those on different scales. The disadvantage here is that if the demand is low, it gives misleading results (Huber et al., 2019). In the light of this information, the performance results from the analysis are as in Table 6.1.

Table 6.1 Forecast Performances based on RMSE, MAPE and MAE

	Train			Test		
Method	**RMSE***	**MAPE****	**MAE***	**RMSE***	**MAPE****	**MAE***
Naive	156.84	0.39	91.76	169.51	0.35	108.72
Snaive	161.34	0.4	95.03	181.09	0.38	117.39
Median	160.64	0.35	112.51	173.07	0.41	124.76
Smedian	116.14	0.33	72.45	152.34	0.38	115.87
Moving Average	129.39	0.37	84.08	168.2	0.38	120.93
LinearReg7	84.61	0.21	63.87	88.9	0.22	62.66
XgBoost7	161.88	0.32	127.09	165.09	0.35	127.66
RandomForest7	**49.18**	**0.1**	**32.54**	98.74	0.23	65.74
LgbmDT7	129.19	0.35	105.04	123.11	0.32	93.48
LinearReg10	83.09	0.19	61.42	**87.23**	**0.21**	**60.48**
XgBoost10	158.36	0.32	127.21	166.19	0.37	130.22
RandomForest10	50.89	0.11	33.63	102.11	0.24	69.45
LgbmDT10	129.19	0.35	105.04	123.11	0.32	93.48
LSTM	148.35	0.32	97.33	149.85	0.38	111.2

*scale-dependent error measure **percentage-based error measure

As can be observed from the table, ML methods give better results than the reference methods. While random forest shows the best performance on the train set in machine learning methods, we can state that this situation returns to the linear regression side on the test set. This implies that random forest overfits the test data.

6.2 Optimization and Cost Performances

After the demand estimation, we examine the separated and integrated optimization results of all methods. In this section, we evaluate the results of the model-based, data-driven, and integrated optimization approaches we mentioned earlier. The reason for using seasonal normalization as a model-based method is to adapt the seasonal distributions to the normal distribution and to strengthen the optimization by increasing the sensitivity. To make proper calculations in optimization, we need to know the overage and underage costs. We obtained this information directly from the company's supply chain manager. Based on this information, we take the ratio of the cost of overage to the cost of underage as 1/3. With this cost information, the target service level is 0.75. We also make percentage cost calculations according to this service level. The cost of the differences is divided by the total actual cost for that period or sample. We can observe the results in Figures 6.1 and 6.2.

Figure 6.1 Cost Performance Analysis-Train Set

Train	**Method**		Target Service Level 0.75	**Method**		Target Service Level 0.75	**Method**		
	Estimation	**Optimization**	**Cost (%)**	**Estimation**	**Optimization**	**Cost (%)**	**Estimation**	**Optimization**	**Cost (%)**
Reference	Naive	S-Norm	44.92%	*ML* LinearReg7	S-Norm	31.14%	*ML* LinearReg10	S-Norm	30.99%
		SAA	50.45%		SAA	36.85%		SAA	35.44%
		QR	42.48%		QR	30.89%		QR	30.30%
	Snaive	S-Norm	42.29%	XgBoost7	S-Norm	83.18%	XgBoost10	S-Norm	82.38%
		SAA	49.98%		SAA	47.09%		SAA	44.84%
		QR	45.48%		QR	26.75%		QR	25.38%
	Median	S-Norm	74.56%	RandomForest7	S-Norm	25.07%	RandomForest10	S-Norm	25.19%
		SAA	63.53%		SAA	19.59%		SAA	20.00%
		QR	62.84%		QR	**15.13%**		QR	15.37%
	Smedian	S-Norm	40.94%	LgbmDT7	S-Norm	59.61%	LgbmDT10	S-Norm	59.61%
		SAA	38.88%		SAA	60.61%		SAA	60.61%
		QR	35.77%		QR	32.76%		QR	32.76%
	Moving Average	S-Norm	40.24%	LSTM	S-Norm	59.65%	Rand7 + Lgbm7	QR	**13.98%**
		SAA	45.99%		SAA	48.69%			
		QR	37.04%		QR	41.42%			

Figure 6.2 Cost Performance Analysis-Test Set

Test	**Method**		Target Service Level 0.75	**Method**		Target Service Level 0.75	**Method**		
	Estimation	**Optimization**	**Cost (%)**	**Estimation**	**Optimization**	**Cost (%)**	**Estimation**	**Optimization**	**Cost (%)**
Reference	Naive	S-Norm	49.22%	*ML* LinearReg7	S-Norm	30.36%	*ML* LinearReg10	S-Norm	29.53%
		SAA	62.21%		SAA	37.76%		SAA	37.83%
		QR	54.62%		QR	30.83%		QR	**29.11%**
	Snaive	S-Norm	55.57%	XgBoost7	S-Norm	91.66%	XgBoost10	S-Norm	93.25%
		SAA	62.05%		SAA	77.20%		SAA	56.69%
		QR	56.91%		QR	46.76%		QR	46.48%
	Median	S-Norm	78.16%	RandomForest7	S-Norm	36.26%	RandomForest10	S-Norm	36.22%
		SAA	71.73%		SAA	40.53%		SAA	44.14%
		QR	79.91%		QR	37.36%		QR	38.77%
	Smedian	S-Norm	58.82%	LgbmDT7	S-Norm	52.44%	LgbmDT10	S-Norm	52.44%
		SAA	58.40%		SAA	52.53%		SAA	52.53%
		QR	59.43%		QR	57.31%		QR	57.31%
	Moving Average	S-Norm	61.65%	LSTM	S-Norm	59.15%	Rand7 + Lgbm7	QR	36.85%
		SAA	68.20%		SAA	63.97%			
		QR	62.77%		QR	64.60%			

Similar to demand forecasting, we can state that machine learning methods generally show superior performance and the costs are lower than the reference methods as expected because the presence of explanatory variables in machine learning methods increases the quality of prediction. Thus, this means that forecast performance has a significant impact on cost performance. Only, XgBoost's performance is so low on the test set because its results below the optimum stock level, and as a result, its cost is multiplied. It is better to repeat that overstock cost is lower than understock because of the company policy. This means that the understock penalty is higher and thus the percentage cost calculations are directly affected. On the other hand, again similar to the forecast metric results, while random forest performance was high in training, it lagged behind linear regression on the test. This shows us that tree-based methods perform below the linear regression method.

When we compare the optimization approaches, we can state that the QR generally shows superior performance. Looking through the test set, we can see that the QR performs on an average 6% better than SAA and S-norm. In the linear regression model where the best result is obtained, QR gives about 8% better than SAA and 0.5% better than S-Norm. As Ban & Rudin (2019), Huber et al. (2019), and Qi et al. (2020) stated in their studies, this is the case that the two-step approaches which means separated demand estimation and optimization multiply the error. This study once again confirms this situation.

7. CONCLUSION

In this study, we consider the data-driven newsvendor problem with a single product and a single period. We approach this perishable inventory problem with the available sales data in two different ways: Sequential demand estimation and optimization and integrated demand estimation and optimization. These approaches allow us to determine the optimal order quantity. We do this to minimize costs and prevent food waste.

Then, the company and data are analyzed. In the methodology, we introduce demand forecasting and optimization. In demand forecasting, we explain machine learning methods such as linear regression, random forest, Xgboost, LgbmDT, LSTM, and ensemble, as well as reference methods such as naive, seasonal naive, median, sessional median, and moving average. In the optimization section, we first introduce the sequential methods, seasonal normalization, and sample average approximation (SAA), and then highlight the integrated methods, quantile regression (QR). Next, in the feature engineering and feature selection section, we explain the features of the variables used in the model and explain how these variables are selected. In the result section, we compare the demand forecast results over MAPE, MAE, and MSE. We also find the costs of machine learning and reference methods according to the objective function and compare these performances.

The key results of the evaluations in the study are as follows: machine learning methods outperform traditional methods in all performance measurements and the integrated demand forecasting and optimization approaches such as QR are more effective than sequential demand forecasting and optimization counterparts. In the demand forecasting phase, we observe that random forest and linear regression methods give more than 10% better results than the best-performing reference method. In the second stage, where optimization approaches are added, we find that QR performs 6% better than other methods.

In the study, we only work on the company's Turkey operations and a single product. In future research, this study can be conducted to cover all operations and based on other products or SKUs. Secondly, we research on a single period in the problem, but a multi-period study may also be the subject of future studies. In addition, working on different companies could be another research area and it also contributes to the reliability of the results obtained from the study. Finally, another future research topic for the study can be extended by working on different machine learning methods or other integrated demand forecasting and optimization approaches.

There are also some limitations to the study. Since we do not access to cost data, we did not have the overage and underage costs. The absence of this cost information does not allow us to optimize and naturally make an inventory decision. Because among the most significant elements for optimization are inventory estimates and overage and underage costs. Since making assumptions about costs will directly affect production costs and profits, we conducted the study by obtaining these costs from the company. The final limitation is that machine learning methods do not consider the importance of the relationship between variables, as they are result-oriented. The lack of this information makes debugging difficult. To avoid this problem, we have eliminated unnecessary variables by applying best-subset selection. Thus, we conducted the study with significant variables and did not need debugging.

BIBLIOGRAPHY

Akkas, A. & Gaur, V. (2021). Om forum — reducing food waste: An operations management research agenda. *SSRN Electronic Journal*.

Akkas, A., Gaur, V., & Simchi-Levi, D. (2018). Drivers of product expiration in consumer packaged goods retailing. *Management Science*.

Akkas, A. & Honhon, D. (2018). Shipment policies for products with fixed shelf lives: Impact on profits and waste. *SSRN Electronic Journal*.

Akkaş, A. & Sahoo, N. (2020). Reducing product expiration by aligning salesforce incentives: A data-driven approach. *Production and Operations Management*, *29*(8), 1992–2009.

Ban, G.-Y. & Rudin, C. (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research*, *67*(1), 90–108.

Belavina, E. (2021). Grocery store density and food waste. *Manufacturing & Service Operations Management*, *23*(1), 1–18.

Belavina, E., Girotra, K., & Kabra, A. (2017). Online grocery retail: Revenue models and environmental impact. *Management Science*, *63*(6), 1781–1799.

Bertsimas, D., Gupta, V., & Kallus, N. (2017). Robust sample average approximation. *Mathematical Programming*, *171*(1-2), 217–282.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Broekmeulen, R. A. & Donselaar, K. H. V. (2019). Quantifying the potential to improve on food waste, freshness and sales for perishables in supermarkets. *International Journal of Production Economics*, *209*, 265–273.

Buisman, M., Haijema, R., & Bloemhof-Ruwaard, J. (2019). Discounting and dynamic shelf life to reduce fresh food waste at retailers. *International Journal of Production Economics*, *209*, 274–284.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Dusoruth, V., Peterson, H. H., & Schmitt, J. (2018). Estimating a local food waste baseline. *Journal of Food Products Marketing*, *24*(5), 654–680.

FAO (2020). Food wastage footprint and climate change.

Hastie, T., Friedman, J., & Tisbshirani, R. (2009). *The Elements of statistical learning*. Springer.

Hu, J., Li, J., & Mehrotra, S. (2019). A data-driven functionally robust approach for simultaneous pricing and order quantity decisions with unknown demand function. *Operations Research*, *67*(6), 1564–1585.

Huber, J., Müller, S., Fleischmann, M., & Stuckenschmidt, H. (2019). A data-driven newsvendor problem: From data to decision. *European Journal of Operational Research*, *278*(3), 904–915.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: with applications in R*. Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kleywegt, A. J., Shapiro, A., & Homem-De-Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, *12*(2), 479–502.

Kostadinov, S. (2018). *Recurrent Neural Networks with Python Quick Start Guide: Sequential Learning and Language Modeling with Tensorflow*. PACKT Publishing Limited.

Kumar, A., Naughton, J., & Patel, J. M. (2015). Learning generalized linear models over normalized data. *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*.

Lee, D. (2012). Turning waste into by-product. *Manufacturing & Service Operations Management*, *14*(1), 115–127.

Levi, R., Perakis, G., & Uichanco, J. (2015). The data-driven newsvendor problem: New bounds and insights. *Operations Research*, *63*(6), 1294–1306.

Levi, R., Roundy, R. O., & Shmoys, D. B. (2007). Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research*, *32*(4), 821–839.

Li, Q., Yu, P., & Wu, X. (2016). Managing perishable inventories in retailing: Replenishment, clearance sales, and segregation. *Operations Research*, *64*(6), 1270–1284.

Natarajan, K., Sim, M., & Uichanco, J. (2018). Asymmetry and ambiguity in newsvendor models. *Management Science*, *64*(7), 3146–3167.

Oroojlooyjadid, A., Snyder, L. V., & Takáč, M. (2020). Applying deep learning to the newsvendor problem. *IISE Transactions*, *52*(4), 444–463.

Papanastasiou, Y. (2020). Newsvendor decisions with two-sided learning. *Management Science*, *66*(11), 5408–5426.

Porteus, E. L. (2002). *Foundations of stochastic inventory theory*. Stanford Univ. Press.

Qi, M., Mak, H., & Shen, Z. M. (2020). Data-driven research in retail operations—a review. *Naval Research Logistics (NRL)*, *67*(8), 595–616.

Reynolds, C., Goucher, L., Quested, T., Bromley, S., Gillick, S., Wells, V. K., Evans, D., Koh, L., Kanyama, A. C., Katzeff, C., & et al. (2019). Review: Consumption-stage food waste reduction interventions – what works and how to design better interventions. *Food Policy*, *83*, 7–27.

Saghafian, S. & Tomlin, B. (2016). The newsvendor under demand ambiguity: Combining data with moment and tail information. *Operations Research*, *64*(1), 167–185.

Scarf, H. E. (1958). *A min-max solution of an inventory problem*. Stanford University Press.

Seubert, F., Stein, N., Taigel, F., & Winkelmann, A. (2020). Making the newsvendor smart - order quantity optimization with anns for a bakery chain. In Anderson, B. B., Thatcher, J., Meservy, R. D., Chudoba, K., Fadel, K. J., & Brown, S. (Eds.), *AMCIS*. Association for Information Systems.

Siegel, A. F. & Wagner, M. R. (2020). Profit estimation error in the newsvendor model under a parametric demand distribution. *Management Science*.

Silver, E. A. & Pyke, D. F. (2017). *Inventory and production management in supply chains*. CRC Press, Taylor & Francis Group.

TISVA (2020). *Sayılarla İsraf Raporu 2018*. Türkiye İsrafı Önleme Vakfı.

Xu, L., Zheng, Y., & Jiang, L. (2021). A robust data-driven approach for the newsvendor problem with nonparametric information. *Manufacturing & Service Operations Management*.

Yang, S., Xiao, Y., & Kuo, Y.-H. (2017). The supply chain design for perishable food with stochastic demand. *Sustainability*, *9*(7), 1195.

Yue, J., Chen, B., & Wang, M.-C. (2006). Expected value of distribution information for the newsvendor problem. *Operations Research*, *54*(6), 1128–1136.