Data Science Methods in Price Prediction

Rianne Hui Yu Weijsters (447964)

Erasmus School of Economics, Erasmus University Rotterdam

MSc Data Science and Marketing Analytics

Supervisor: Prof. dr. P.H.B.F. Franses

Second assessor: dr. V. Avagyan

Supervisor PwC: dr. M.O. de Kok

August 16, 2021

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

This study explored different data science methods in contribution margin prediction by analyzing historical sales data from a Business-to-Business chemical company. The aim of this study is to shed light on different predictive modelling techniques which aids marketers in developing a price guidance that will lead to profit maximization. The following methods were assessed on different data samples using the RMSE and R^2 as performance metrics: multiple linear regression, stepwise forward- and backward selection, elastic net, random forest, k-nearest neighbors, and neural network. The three data samples that were investigated included the entire dataset, subset 1 excluding categories with less than 5% of the total number of observations, and subset 2 containing only observations within the interquartile range. The results demonstrated that all models performed best on subset 1. Additionally, evidence suggests that the random forest model obtained superior performance (RMSE=.311, R^2=.626) compared to the other models. The findings indicate that in general prediction estimation improves when all categories are well represented by the data. Further research is necessary to understand if the results are applicable to other industries and business sectors as well.

Keywords: pricing, machine learning, multiple linear regression, stepwise selection, elastic net, random forest, k-nearest neighbors, neural network

Contents

Introduction

The rise of new, innovative, and more advanced technologies is disrupting the way businesses operate. The rapid advancement in stronger computational power has resulted in a paradigm shift that has transformed the marketing landscape to a more data-driven approach. This phenomenon is often conceptualized by the popular metaphor stating that data is the new oil of the economy, and analytics is the combustion engine (Sondergaard, 2011). Data science methods, including Machine Learning (ML), have altered the nature of managerial decision-making, especially when it comes to the marketing strategy of businesses. The abundance of data and the adoption of advanced statistical methods and ML techniques have transformed the way of working for marketing practitioners to a more technology-enabled way. It is therefore not surprising that researchers recently called for research in this relatively new area in the marketing domain (see Kumar et al., 2021; Shah & Murthi, 2021, for more detail).

For decades, one of the most traditional frameworks adopted in the field of marketing has been the marketing mix, also referred to as the four Ps – product, place, price, and promotion – which is a widely recognized and well-established framework that businesses commonly use as a foundation for their decision-making to pursue their marketing strategies and to achieve their target goals (Borden, 1964). Price is one of the elements of the marketing mix and is often perceived as the most fundamental one when it comes to managerial decision-making. While product, place, and promotion create added value, such as benefits, to the customer, price rather “harvests” the value created by the other elements of the marketing mix (Schindler, 2011). Price is the only P that yields direct revenue and serves as an indication of the value and position of a business.

Multiple scholars have stressed that deploying a superior pricing strategy leads to a competitive advantage (Baker et al., 2010). Previous research on pricing theory has shown that price is one of the most fundamental profit levers for businesses. For instance, Baker et al.

(2010) concluded based on a study across 1200 companies worldwide, on average a 1% increase in price leads to an 11% increase in profit, while a 1% increase in sales volume will, on average, increase profits by 3.7%, ceteris paribus. In other words, small price modifications can have a big impact on profit, hence huge potential gains can be realized for businesses. In addition, also industry experts from the field itself address the importance of pricing. For example, Warren Buffett, the billionaire chairman and CEO of Berkshire Hathaway:

> The single most important decision in evaluating a business is pricing power. If you've got the power to raise prices without losing business to a competitor, you've got a very good business. And if you have to have a prayer session before raising the price by 10 percent, then you've got a terrible business. (Buffett, 2011)

Yet, managing price successfully is a challenging task that can, when managed on insufficient grounds, do more harm than good to the value of a business. A marketer's goal is to set prices in such a way that profit leverage reaches its maximum. There exists a considerable body of literature on different pricing strategies that marketers can use to sell a product or service. P strategies typically fall into one of the following three categories: cost-based pricing, competition-based pricing, or value-based pricing (Schindler, 2011). The question that then needs to be addressed entails: how will the final market price be determined? According to the laws of economics, in perfectly operating markets, price is determined at the equilibrium where supply meets demand. In practice, however, deciding on a pricing strategy is seldomly a simple process. Pricing is rather complex in nature as it is linked to different aspects of a business, such as its core values and key objectives, and usually involves more than one strategy to not lose out on potential profit. A common approach for businesses to evaluate the price of a product or service, is the use of a pricing framework that offers marketers guidance from price strategy to execution. An example of such a framework is PwC's Conceptual Pricing

Framework as illustrated in Figure 1 in the literature review section, which will then be further elaborated on (PricewaterhouseCoopers, 2021).

As mentioned, in the past several decades the marketing mix 4Ps framework has played an important role in guiding marketers in their decision-making. A typical drawback of the 4Ps framework is that it does not take customer's value perceptions into consideration. There is, however, a growing appeal for a framework that does not limit businesses in reaching their full potential, especially for those operating in the Business-to-Business (B2B) sector. For example, Ettenson et al. (2013) introduced an adapted version of the marketing mix specifically designed for B2B businesses, namely SAVE: Solutions, Access, Value, and Education. For the price element, in specific, the emphasis has shifted to a more value-based approach in which marketers focus on the benefits and solutions that the product or service brings to the customer, rather than focusing on the costs and its profit margin.

Despite the introduction of a different marketing mix for the B2B sector, the body of literature on this new approach remains rather scarce. More specifically, no study to date has examined the process of determining the right price utilizing data that will optimize profit for B2B businesses using the SAVE framework. Especially in today's digital realm in which technology continuously advances, researchers stress the urgent need for a more in-depth understanding of how data science techniques, such as ML, impact the marketing tools used to enhance decision-making (Kumar et al., 2021).

Therefore, this thesis aims to explore which ML methods can be applied to make pricing decision-making in B2B businesses more data-driven, which will subsequently help marketers in getting the right price that will yield the highest profit. Specifically, this thesis focuses on how historical sales data can be used to create a data-driven price guidance for sales reps in B2B businesses. Based on this, the following research question has been derived:

"*Which Machine Learning Technique is Most Effective for Optimal Price Prediction for B2B Businesses?*"

This thesis aims to examine this research question by investigating historical transaction data of a B2B company. In doing so it explores different data science methods and thereafter performs model assessment. The goal of this thesis is to contribute to the small body of pricing literature focused on pricing analytics, by providing a critical viewpoint on which model performs statistically best, but also on which model could be most effective from a practical point of view. The reason why this is relevant is that a more granular and differentiated price guidance allows for informing sales reps more precisely on the price opportunity across different market groups. Consequently, chances of winning deals at the optimal price are increased, and this creates an upward push on the price, and thus on the value of the business. This can be achieved by ultimately informing sales reps what bottom, target, and best possible prices are, based on what the business was able to charge similar customers for similar products in the recent past. More precisely, based on an accurate estimation, pricing practitioners may answer the following three questions with higher certainty:

(1) "What price drivers can be identified?"

(2) "What segments on both product and customer characteristics can be identified?"

(3) "What price guidance can be derived, optimized for those price drivers and segments?"

To gain a better understanding of the three questions noted above, first it must be examined which data science technique is most accurate in predicting the price. This thesis will do so by investigating the following sub-questions:

(1) "What B2B pricing data science methods exist?"

(2) "What data science techniques are relevant for price prediction?"

(3) "How do the methods perform, and which method has the highest performance in estimating the price?"

By investigating the research and sub-questions described above, the present thesis aims to deliver several key contributions to the field of marketing, more specifically pricing, with a substantive focus on data-driven marketing. First, the main objective of this study is to shed light on the use of a ML techniques on pricing. By doing so, it aims to investigate which methods may best lead to price determination. Up to date, this approach has not been touched upon by pricing academics. Second, the objective of this study is to bridge the gap between academics and pricing practitioners by investigating a practical business case of the B2B sector, which, to my knowledge, has not been addressed in academic literature yet.

This thesis starts with a review of the relevant literature, touching upon the 1) evolution of marketing to a data-driven transformation, 2) customer value-based pricing, and 3) pricing frameworks exploited in B2B businesses. Following the literature section, a description of the data and the data preparation steps is given. Next, the relevant methods used to investigate the research question is explained in the method section. After that, the statistical findings are reported in the results section. Furthermore, a comprehensive discussion on the findings is followed in the discussion section, in which this thesis aims to answer the research question, shed light on the managerial implications, and provide the research limitations and recommendations for future research. Last but not least, the final section entails the conclusion of the present research.

Literature Review

To further explore the role of data science in price prediction the present study will start by defining what pricing entails and by looking into the development of how pricing has started to grow into an independent business domain. After this, current literature on data-driven marketing and data-driven pricing will be explored. More specifically, the shift from product-based marketing to customer-based marketing, and the shift from cost-based pricing to value-based pricing will be discussed to gain a better understanding of the development of data-driven marketing. The final section will bridge the gap between data-driven marketing and value-based pricing by explaining how these trends are connected with customer segmentation.

Pricing

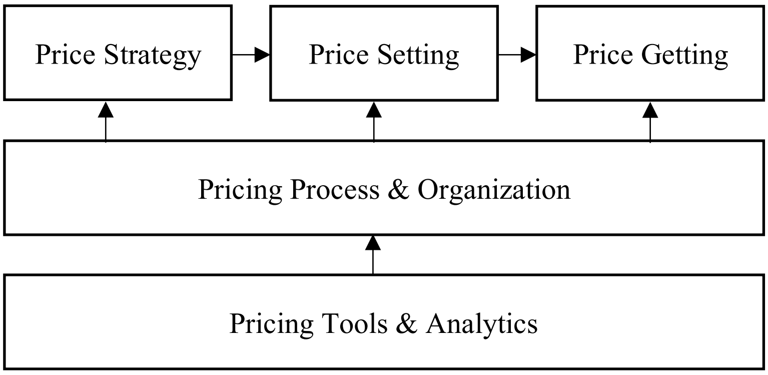
What is Pricing?

Price stems from the Latin word pretium, meaning worth or value, and is typically seen as a certain amount of money a product or service costs. According to Lancioni et al. (2005), pricing can be viewed from two different standpoints. On the one hand, from the customer's point of view, the price can be defined as the value a business brings to its customers. Customers pay a certain amount for a product or service which is equivalent to the quantification of the added value the customer receives by consuming what has been bought. On the other hand, from a business perspective, price is rather defined as a strategic decision that communicates a value for which a business sets itself on the market and is used to realize profits. The process of developing a pricing strategy, however, remains a complex task that is often not yet fully utilized or well-understood by businesses (Baker et al., 2010). A reason for this could be explained by the findings of a study conducted by Hinterhuber and Liozu (2012), where it was concluded that, regardless of the industry or country the business is operating in, pricing is considered as a difficult skill one must acquire to successfully perform the two fundamental components of pricing: price setting and price getting. It is important to note that multiple academics have indicated that pricing is an extremely difficult task, as a full understanding of the elements involved in pricing is required before it may lead to a superior business position (Ingenbleek et al., 2003). Hence, this may explain why in practice the majority of businesses do not utilize the advantages of pricing to their full potential yet.

The overall pricing process is conceptualized in Figure 1, which is a common framework pricing practitioners and academics use as a foundation for pricing (Baker et al., 2010; Hinterhuber & Liozu, 2012). As can be seen in Figure 1, first the price strategy must be decided on, after which price setting and price getting can be performed. The overall process is enabled and underpinned by two elements, namely the deployment of a central pricing organization, and the utilization of pricing tools, such as software solutions, to successfully improve monetization. The price strategy must be considered first and is often done by evaluating the pricing goals and objectives, e.g., profit or market share. In addition, the customer, and the channel segmentation the business is operating in, and the price and brand positioning compared to competitors, need to be examined to gain a thorough understanding of the product or service on sale. On top of that, the value or benefits the product or service brings to the customers, that is its value proposition, needs to be well-understood before the right pricing strategy can be chosen (Hinterhuber & Liozu, 2012).

Figure 1

Conceptual Pricing Framework



Note. Adapted from *PricewaterhouseCoopers*, 2021.

Once the pricing strategy is set, the second element of the pricing process involves price setting. According to Hinterhuber and Liozu (2012), price setting is referred to as the process of deploying different approaches to decide on the list prices, also known as the recommended or suggested prices, and discount structure of a company's products and services. In other words, the goal of this stage is to establish the right price. Price setting practitioners commonly consider the trade-off between capturing a customer's willingness to pay (WTP) and the costs of a product. In addition to that, price setting also involves the evaluation of other external factors such as market demand and dynamics, competitor's behavior, and economic trends. Price setting aids marketers to spot pricing opportunities early in time. Typically, the foundation of such pricing approaches is either cost-based pricing, competition-based pricing, or value-based pricing (Hinterhuber and Liozu, 2012; Schindler, 2011). These approaches will be further elaborated on in the Customer Focused Marketing and Value-Based Pricing section below.

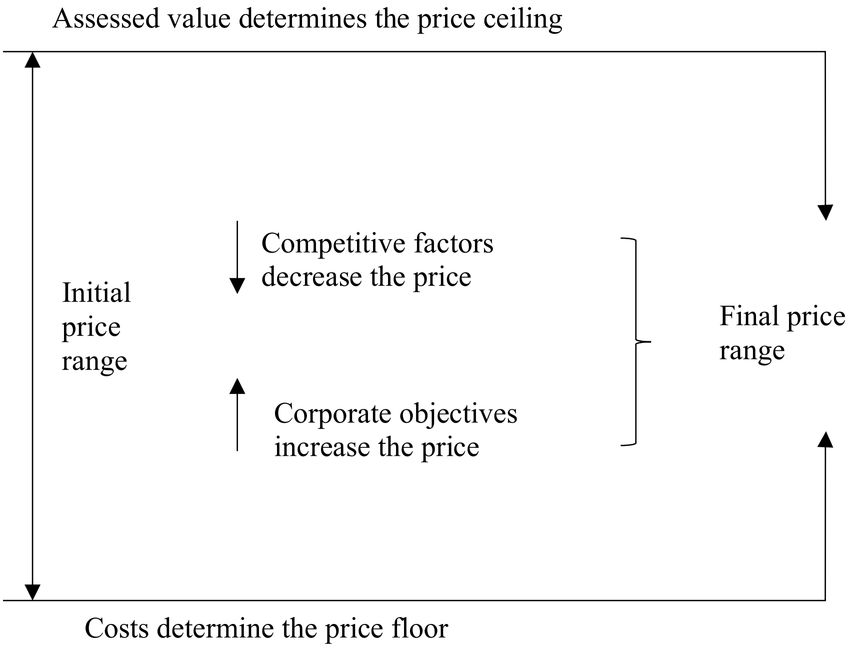
Next to price setting, the authors note that price getting relates to the steps a business undertakes to get a fair price for its products and services in the market: it involves the implementation of the pricing approach set in the price setting stage. A key element here is providing price guidance to the sales team of the business, mainly for those operating in the B2B sector. Other important elements that are involved in the price getting process are, for example, value selling and value stories, and deal reviews for B2B businesses mainly. In the case of B2C businesses, factors such as discounts and promotions management, and behavioral tactics play a more fundamental role. A price guidance creates price recommendations based on, for instance, market and competitor insights, or historically realized transactions, and therefore helps marketers in getting the right current price for their products or services (Salesforce, 2020). To summarize, the price getting stage is a crucial stage for companies to realize their target goals, i.e., the set price, and to create winning deals.

The Process

As mentioned above, pricing is a complex task and consequently, academics imply that it is a skill one must acquire to perform pricing successfully. To illustrate the complexity behind pricing further, Figure 2 provides a demonstration of the common elements involved in deciding on the price (Ingenbleek et al., 2003). According to this framework, typically, price ranges between the costs of the product or service, and the customer's WTP, which in turn depends on the customer's value perception of the product or service being offered, and this price range is also known as the initial price discretion. The highest price that businesses can set, which is equivalent to the price ceiling, depends on the customer's maximum WTP. Conversely, the lowest possible price, namely to the price floor, depends on the variable costs of the product or service. To determine the final price, factors such as market competition, and corporate objectives have a strong impact on the price ceiling and price floor, respectively. To clarify this further, high competition may lead to lower WTP, which reduces customer's value perceptions, which in turn results in a decrease in the price ceiling. Conversely, corporate objectives, such as a target margin, increase the price floor (Ingenbleek et al., 2003). To summarize, it can be concluded that pricing is a complex task that may depend on factors that are hard to quantify, such as customer value perception. Nevertheless, this indicates that it is extremely important that pricing practitioners obtain a thorough understanding of customer's value perceptions, by for example conducting conjoint analyses, before the right price can be determined.

Figure 2

The Price Determination Process Framework


Assessed value determines the price ceiling
Competitive factors decrease the price
Initial price range
Final price range
Corporate objectives increase the price
Costs determine the price floor

Pricing as a Business Domain

For a long time, business managers have seemed to neglect the advantages that pricing brings to businesses (Baker et al., 2010). Traditionally, price was seen as a part of the marketing mix, together with product, place, and promotion, where price was the only element that brought in the value created by the other marketing mix elements to a business (Schindler, 2011). More recently, however, the body of literature on pricing has tremendously expanded and gradually researchers and managers start to recognize the added value and opportunities pricing itself brings to businesses (Liozu et al., 2012). On the one hand, the awareness of price opportunities may lead to profit maximization if the right price is set, but on the other hand, this also means that selling your products and services for the wrong price will result in a loss in profits (Baker et al., 2010). Consequently, due to its large impact on businesses, research on this relatively new domain is substantively growing, and within businesses more and more resources are invested to pricing related issues. Since businesses pay more attention to the benefits that pricing brings, pricing is slowly starting to evolve as an independent business domain (Hinterhuber & Liozu, 2012). As a result, businesses will gradually become better skilled in the process of price setting and price getting, and hence take more advantage of the opportunities that pricing brings.

The need for an independent pricing domain is slowly being recognized, and according to academics the success of those pricing domains is strongly supported by the growth in analytical capabilities which enable the process of setting and getting the right price (Shah & Murthi, 2021). In line with this, Baker et al. (2010) suggest that the reason why pricing is not fully exploited by businesses yet is due to the large amounts of current and sophisticated customer data that is necessary to perform pricing strategies. The availability and usability of such data have, however, improved tremendously during the past decades due to technological developments (Kumar et al., 2021). Moreover, researchers suggest that the recognition and deployment of sophisticated analytics in businesses is a way to gain a competitive advantage in today's economy (Davenport, 2006). Altogether, although a few scholars have illuminated the power pricing entails, currently a minority of businesses realize its added value, and hence benefit from it. Yet, awareness of it is growing, resulting in more skilled pricing practitioners over time.

Data-driven Marketing and Pricing

The evolution of traditional product-centric marketing to data-driven customer-centric marketing is highlighted by Shah and Murthi (2021) more substantively in five consecutive stages: 1) creativity, 2) relevancy, 3) analytics capability, 4) accountability, and 5) technology. The first stage of creativity was primarily present in the late 19th century, which was when the scope of marketing mainly involved the communication of the product or service to the customer by means of promotions and advertisements. Therefore, creativity was a vital aspect of marketing to successfully promote the product or service, and thus marketers conducted quantitative surveys and experiments to improve their communication channels. Gradually, marketers entered the second stage of customer relevancy, where the need for customer data expanded as customer-centricity became more relevant, and marketers started to put themselves in the shoes of their customers to successfully serve customer's wants and needs (Kumar et al., 2006). Next, the third stage represents the analytical capabilities that evolved at the end of the 20th century, when big data gradually emerged, and marketing analytics continuously developed to support marketer's decision-making with the main goal of profit maximization. Examples of such techniques are (choice)-based conjoint analyses to model consumer choices and Bayesian methods for consumer behavior modeling (Wedel and Kannan, 2016).

After marketers deployed analytical methods in their decision-making, they entered the fourth stage of accountability in the twenty-first century. The main focus of this stage was to ensure that the financial accountability that marketing analytics entailed, was recognized by managerial decision-makers. A series of recent studies suggest that the impact of advanced analytics, and its further developments, on the marketing mix is currently still highly promising. The reason for this is that more accurate market insights can be gained which leads to better decision making, resulting in optimal prices, which in turn improves profit maximization (Shah & Shay, 2019; Shah & Murthi, 2021). Consequently, the accountability of marketing analytics in businesses received a more profound role. More recently, businesses entered the final stage, which the authors named new technologies, where more advanced methods, such as artificial intelligence, have become a foundational pillar of marketing analytics, where especially the profound insights resulting from the predictable power of advanced analytics and ML methods creates added value and support to businesses (Wedel & Kannan, 2016).

Admittedly, research on this area of new technologies, remains rather scarce due to the recent emerging trend (Wedel & Kannan, 2016). No attempt has been made to investigate what data science methods could be used to optimize marketing decision-making, such as pricing problems. Although the implications of a price guidance based on predictive modelling have been recognized, up to date scholars have not considered which method may be suitable for pricing related issues. Nevertheless, Wedel and Kannan (2016) point out the need for research on the application of ML methods and other advanced analytical techniques in the marketing domain for both the B2C and the B2B sectors. The reason for this is that there is an increased demand for analytical methods that can cope with large amounts of data, i.e., big data, and with the challenges arising from the continuously changing environment as well as increased customer demands. This is further supported by a more recent study conducted by Shah and Shay (2019) who shed light on the impact of analytical capabilities on marketing. The researchers argue that in business applications, ML is often used for, amongst other purposes, pricing strategies. Typically, methods such as linear regression, decision trees, and random forests are commonly applied techniques to understand the price-drivers of products and services (Chui, 2018). The effectiveness of different statistical methods however remains unaddressed, and it can therefore be concluded that this particular topic on what methods are employed in pricing is yet briefly addressed in academic literature. To fill in this gap, an urging question that needs to be addressed therefore entails: what data science methods can pricing practitioners employ to set the right price that allows for profit maximization. This thesis aims to explore this question further in the method section by analyzing historical customer data.

Customer Focused Marketing and Value Based Pricing

Numerous studies have addressed how the adoption of technological advancements in society has influenced and changed what marketing entails – see for example Kumar (2015), Kumar et al. (2021), Wedel and Kannan (2016), and Shah and Murthi (2021) to name a few. Traditionally, the main focus of marketing was held in the exchanges of goods and services with the main objective to sell products, meaning that marketers exploited a product-centric approach (Shah et al., 2006). Nowadays, however, the focus has shifted to a more value-based approach in which the central idea lies around delivering value to the customer through customer relationship management (CRM), and by acting according to customer’s needs, while keeping a long-term perspective in mind (Sheth & Uslay, 2007). The shift to a more value-driven approach is an active response to the augmentation of data utilized by businesses, and hence changed the nature and scope of marketing. More specifically, researchers argue that the wide integration of the internet in society, in combination with advanced analytics, has driven marketers to transform to a customer-oriented approach that allows for customer value creation, beyond the value created from just selling, in which responding to customer wants and needs has become the key concept of marketing (Kumar, 2015). Examples of evolved ways of value creation include opportunities in, for example, customization. Besides, according to Sheth and

Uslay (2007), the new role of marketing in businesses is applicable in both the B2C and the B2B sector, mainly due to the wide adoption and integration of the use of technology and access to data across all businesses, sectors, and industries. Along similar lines, Kumar et al. (2021) adds to this by suggesting that greater improvements in data management and storage capacity will drive B2B businesses to exploit customer relationships more and more to develop a customer-focused strategy.

Likewise, pricing has also become more driven by customer value perceptions (Hinterhuber, 2008). Businesses commonly employ one of the earlier mentioned approaches, namely cost-based or competition-based strategy, that lay the foundation for pricing. However, more recently, a shift from cost- or competition-based pricing to value-based pricing has been noticed (Hinterhuber & Liozu, 2012). The remaining of this section will further elaborate on each of these approaches.

First, businesses deploying a cost-based pricing strategy set prices based on the actual costs and a preselected margin of profit. An advantage of cost-based pricing is that when production efficiency increases over time, the costs per product will decrease. Hinterhuber and Liozu (2012) argue that a downside of employing a cost-based strategy, however, is that businesses exploiting a cost-based pricing strategy are prone to lose on potential profits when competitors become more cost-efficient or when prices are not set optimally to capture customer’s WTP. According to the authors, a big advantage, however, is that cost-based pricing is an achievable strategy for many businesses, because no thorough understanding of the customers or market is needed. Therefore, many B2B businesses employ such a strategy, due to the limited availability of customer data. Second, competition-based pricing is an approach in which businesses base their prices on how competitors price their products or services. Similar to cost-based pricing, customer’s WTP is not taken into consideration, hence businesses following this strategy will lose out on profits. Nevertheless, according to Liozu et al. (2012), competition-based pricing is still the most adopted strategy in pricing. This could be explained by the fact that like cost-based pricing, no in-depth market knowledge is needed, and therefore competition-based pricing is fairly easy to implement, providing that competitor information is available.

The third pricing approach is customer value-based pricing, which is often considered as the most advanced approach as it is based on customer's value perception and their WTP for the product or service. A critical question here is how marketers and sales teams can create additional customer value to increase the WTP of customers. To do so successfully, a full comprehension of customer's needs, wants, value perceptions, market segments, and WTP is necessary to base decisions on. This is where data and analytical capabilities play a substantive role (Hinterhuber & Liozu, 2012). Customer value-based pricing is becoming increasingly popular amongst scholars and businesses because this approach allows for profit maximization. Evidently, previous studies have shown that value-based pricing leads to profit maximization, because a more comprehensive judgment of the WTP can be made if the historical customer data is available (Ingenbleek et al., 2003). In addition, in a pricing literature review, Ingenbleek (2007) suggests that the customer value is determined based on a trade-off the customer makes between how much value he or she perceives the product or service to be, and what the customer is willing to sacrifice.

Despite the promises this strategy brings, researchers note that just a few businesses adopt this strategy as this approach is not yet well understood by marketeers (Liozu et al., 2012). Consequently, businesses may suffer from serious shortcomings. This phenomenon is clearly highlighted by Hinterhuber's (2008) study based on 81 businesses operating on a worldwide scale in both the B2B and B2C sector, where it was concluded that 80% of the businesses investigated, fell back to cost-based or competition-based pricing when practicing value-based pricing. It was found that the reasons for this failure could possibly be explained by difficulties businesses often encounter, such as deficits in value assessment for different market segments. Liozu et al. (2012) adds to this, by stating that businesses who follow a competition- or cost-based strategy appear to have insufficient understanding of the tools and skills needed to successfully execute value-based pricing.

Segmentation

To bridge the gap between data-driven marketing and value-based pricing is where segmentation comes into place. Incorporating segmentation in a pricing strategy enables the process of customizing prices according to different segment levels or groups of customers (Bouter, 2013). Segmenting based on customers is equivalent to identifying different WTPs on which prices will be subsequently based. The main objective of this approach is to group customers based on similar wants and needs to enable a targeted approach (Martin, 2011). A wide range of literature indicates that segmenting customers based on value perceptions improves CRM (Kim, 2006). As a result, customer loyalty increases, which ultimately increases the value of a business.

To practice value-based pricing, however, a solid understanding of customers is of crucial importance (Hinterhuber, 2008). This indicates that a substantive amount of customer data is needed to successfully perform segmentation. More precisely, when historical customer data is available, from which the customer value perception can be estimated, sellers can match prices according to customer's highest value perceptions of the product or service being sold, i.e., match prices according to customer's WTP (Ingenbleek, 2003). Academic pricing literature shows that there are multiple ways to measure customer's WTP, such as analyzing real purchase data, or conducting choice-based conjoint analysis (Miller et al., 2011). The authors, however, argue that defining customer's WTP by employing such measures leads to hypothetical bias as one can only make an estimation. Despite the possible bias involved, evidence shows that the estimated WTP may still lead to the right price setting (Miller et al., 2011). According to Bouter (2013), charging prices according to customer value perceptions is especially impactful in cases when the WTP ranges on a large scale and the variable costs are low. This particular situation is also clarified by the Price Determination Process Framework of Ingenbleek et al. (2003) shown earlier in Figure 2. The phenomenon of successful price customization is also referred to as price fencing in academic literature (Bouter, 2013). A key question here is to investigate whether the market segments are large enough and if they have a steady position in the market i.e., will not decrease in size, to earn profit. It is, therefore, critical for pricing practitioners to gain a full understanding of the market they serve before segmentation is done successfully, and a boost in profits is obtained (Martin, 2011).

Next to segmentation based on customer needs and wants, there are several other ways on which segmentation can be applied, such as taking into account the different purchasing channels or volume of the order (Bouter, 2013; Martin, 2011; Baker et al., 2010). Grouping customers based on the purchasing process involves factors such as geographic location or time of purchase. Moreover, a different way of segmentation includes product and service level. Basing price on product type and service, also known as product versioning, means constructing multiple product variations to meet the needs of customer segments with differing WTP. The aim of this form of segmentation is to obtain a tiered pricing model which gets customers to sort themselves into different groups based on their differing value perception and WTP (Baker et al., 2010).

Given the review of current pricing literature above, it can be concluded that customer research is of critical importance to decide at what level to price a product or service for different customer segments. Only if customers have different WTP, segmentation allows for price differentiation and enables the process of customer-focused marketing and value-based pricing, where prices are determined by customer personas that represent different customer wants and needs. Consequently, the objective of pricing i.e., profit maximization, can be achieved. Yet, to successfully implement a market segmentation, businesses need to obtain a comprehensive market understanding. More importantly, if the goal is to segment on customer's wants and needs, pricing practitioners must know customer's WTP. Therefore, especially in the latter case, a large amount of customer data is needed, together with strong analytical capabilities that allow pricing practitioners to set the right price for the right customer (Shah & Murthi, 2021).

Data

The present study aimed to investigate which data science method performed best at predicting the price. To evaluate different data science techniques, the study used historical transaction data from a global leading chemical producer. The dataset contained in total 42.386 historical sales transactions which took place between 2017 and 2020. All sales transactions were categorized according to their application purpose, of which in total 94 different types of applications were present. For the purpose of this study, a subset of the application that generated the highest revenue was created, which contained a total of 14.797 observations (see Figure 1, Appendix).

The data was first explored and pre-processed for analysis. As for the data cleaning steps, the data was checked for missing values, duplicates, outliers, and negative transactions. Due to the large amount of data, observations with missing values were completely removed. In addition, no duplicates were found. Furthermore, one outlier was detected, which was subsequently removed, as it was assumed that this data point was an incorrect data entry due to its extremely unusual value compared to the rest of the data. Last but not least, transactions that were returned contained a negative revenue, and hence were removed. The remaining 14.776 observations were left for analysis.

The dataset consisted of a total of 21 variables. Variables with no meaningful information, such as the material number and customer name, were removed for this study. The present study used the contribution margin, i.e., price minus variable costs, rather than price as the dependent variable, because the raw material costs fluctuated strongly depending on different marketing conditions, yet the contribution margin remained constant. As no additional data on the market conditions of the chemical industry was available to combat this problem, for the purpose of the study it was decided to use the contribution margin as the dependent variable. Moreover, profit maximization is achieved for prices with the highest contribution margin. The original contribution margin variable was divided by the volume of the transaction so that the contribution margin represented the margin in euros per kilogram.

Regarding variable selection, the time when the transaction took place was not taken into consideration for the analysis, as it was assumed that the contribution margin was not affected in the short term. Moreover, a new variable namely total annual customer volume was added to make up for the size of the customer. After removing all irrelevant variables, 10 variables remained for analysis. The following remaining 9 variables were used as independent variables: material type (six categories), polymer (17 categories), additive (32 categories), filler (24 categories), color (6 categories), segment (8 categories), customer classification (9 categories), country (48 categories), and total customer volume. A summary of the variable definitions and the descriptive statistics can be found in Table 1 (a-c) and 2 respectively in the Appendix.

For the final data preparation, first one-hot encoding was performed on all categorical variables to obtain all dummy variables, where a 0 indicated that the category was not present for a specific transaction, whereas a 1 indicated that the category was present for that transaction. After one-hot encoding was applied, the data consisted of 133 variables. Next, the dataset was split into a train (10344 observations) set and test (4432 observations) set, for 70 and 30 percent, respectively. In the final step of the data preparation, the distribution of the dependent variable was checked, and due to the strongly skewed distribution, a log transformation was done to make the variable more normally distributed, and hence to improve model fit (see Figure 2 and 3, Appendix).

Method

As became clear in the literature review, pricing modeling techniques have been barely addressed in academic literature. Therefore, this study aims to evaluate the performance of multiple predictive modeling techniques. More specifically, the present study investigates both parametric methods and non-parametric methods on different subsets of the data. The reason for studying both parametric and non-parametric methods was to explore whether more flexible approaches would be preferred over non-flexible approaches that make prior assumptions about the data. For the analysis, the following six models were investigated: multiple linear regression (MLR), stepwise forward- and backward regression (SW), elastic net (EN), random forest (RF), k-nearest neighbors (KNN), and neural network (NN). The reason for selecting linear regression models is that it serves as the foundation of statistical modeling. Besides, linear regression is perceived as a simple approach to explain a relationship between the response and explanatory variables. This is especially important for pricing practitioners who may lack knowledge of statistical modeling. In addition to linear regression, stepwise linear regression and penalized linear regression are considered as these are extensions of the MLR, with the aim to improve estimation accuracy and interpretation (James et al., 2013). On top of that, the non-parametric methods, RF, KNN, NN, were also assessed as these approaches are known to be more flexible in their predictions as they do not make prior assumptions about the distribution of the data. All six models were evaluated on three data samples to gain a better understanding of why certain models perform better than others. The samples on which the models were evaluated are summarized in Table 1 shown below. First, the models were conducted on the entire data set, including all observations. Second, a data subset was created which removed categories that entailed less than 5% of the total observations. By investigating this subset, this study aims to examine whether small categories will affect prediction estimation. Third, another subset was created which included only observations that were within the interquartile range. The analyses of this study were conducted using the statistical software RStudio.

Table 1

Summary of the Data Samples

Data Sample
Entire dataset
Subset 1 – categories entailing less than 5% of the total observations are excluded
Subset 2 – only observations that were present within the interquartile range

First, each model was trained on the training sample, and subsequently, the model performance was assessed using the test sample. For the purpose of this study, the performance on the test set was of interest, because it is relevant for pricing practitioners to know which model performs best on unseen data for future cases. Nevertheless, the performance of the train set and the test set were compared to investigate if overfitting was present. In such a case, the model would be too well trained on the train data, but would not perform well on the unseen test data. With regards to the metrics used to assess the model's performance, the root mean squared error (RMSE) and the R-squared (R^2) were used to evaluate and compare the accuracy of each model. By focusing on these two performance metrics, all approaches could be easily compared with each other. The rationale to consider these particular performance metrics is based on James et al. (2013) who states that these metrics are commonly used to assess methods with a quantitative response variable. This is also confirmed by a more recent research conducted by Chicco et al. (2021). The author adds to this by concluding that the R^2 is a better performance metric compared to the mean square error, the absolute mean error, or the mean absolute percentage error, in particular for regression models. Moreover, Chai et al. (2014) demonstrated that using the RMSE as a performance metric is better than the absolute mean error when the model error follows a normal distribution. In addition, according to academic literature, no single universal performance metric exists. Hence, typically a combination of different metrics is used for model assessment (Chai et al., 2014).

The RMSE is given by equation 1:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y_i} - y_i)^2}{n}} \tag{1}$$

where $\hat{y_i}$ is the predicted value for observation i, and y_i is the actual value for observation i. The RMSE, also known as the test error, can be defined as an estimate of the true error, which cannot be computed due to the unknown data, and hence the test error is often used as an alternative metric to evaluate the model performance (Ghatak, 2017). The lowest RMSE across the models was looked for, as it would indicate how close the predicted value $\hat{y_i}$ is to the true value y_i (James et al., 2013). In addition to the RMSE, the R^2-statistic is also a common metric to assess the performance of a model, because the R^2-statistic indicates how much variation is explained by the model (James et al., 2013). A high R^2-statistic was looked for, indicating that the model was able to explain a large proportion of the variability in the dependent variable. In the following subsections, each method will be explained in more detail.

Multiple Linear Regression

Linear regression is a supervised learning technique that is easy to interpret and tends to perform well on real data problems (James et al., 2013). Linear regression aims to fit a model to the train data by applying the least squares procedure. In other words, a model is fitted by minimizing the residual sum of squares (RSS), e.g., minimizing the difference between the predicted value and the true value for all observations as depicted by the following equation (2):

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 \tag{2}$$

where y_i is the observed value for observation i, β_0 is the intercept, β_j is the estimated coefficient for predictor j, and X_{ij} is the predictor for observation i. Linear regression can be defined as a predictive modeling technique with an input vector $X^T = (X_1, X_2, \dots, X_p)$, and an output prediction Y, where the input vector X^T contains the values for the dependent variable T. The linear regression can therefore be rewritten as equation 3:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j\beta_j \tag{3}$$

where β_0 is the intercept, β_j is the coefficient for transaction j, and X_j is the independent variable for transaction j. Hence, to illustrate, given the p dependent variables, the MLR model of the present study can be described as (4):

$$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \cdots + \widehat{\beta_p}X_p + \epsilon \tag{4}$$

where Y is the contribution margin in euros per kilograms, $\widehat{\beta_0}$ is the intercept, $\widehat{\beta_p} = (\widehat{\beta_0}, \cdots, \widehat{\beta_p})$ is a vector of the predicted coefficients of the dependent variable X_p, and ϵ is the error term that captures everything that is not explained by the model (James et al., 2013; Zou & Hastie, 2003).

Stepwise Forward Backward Selection

Despite the simplicity of linear regression, it has also received some criticism as it often fails to accurately capture the true relationship between the predictors and the response variable (Zou & Hastie, 2003). In other words, the model fit, and interpretation of linear regression may not yield useful results. Therefore, to gain a better understanding of which dependent variables are relevant for explaining the contribution margin, variable selection can be performed as an alternative approach, especially in cases where the number of predictors is large. The reason for this is that with a large number of dependent variables, it is highly likely that, by chance, one of the dependent variables will appear to have a significant effect on the independent variable, even if no true relationship exists (James et al., 2013). An efficient solution to this is to perform a variable selection method. The most common variable selection methods are forward selection, backward selection, and mixed selection which combines the principles of both forward and backward selection. The feature selection approach aims to improve the accuracy, and thus prediction performance, of the MLR model by removing irrelevant variables.

SW regression employs an algorithm that can be explained by the following steps. First, M_0 serves as a null model with no predictors and is used as the starting point. Second, predictors are iteratively added and removed depending on whether the fit of the model improved. As a result, a model without redundant predictors would be obtained. Hence, by following this approach, it could be investigated which predictors would have an actual effect on the contribution margin. The Akaike Information Criterion (AIC), was used to assess the model performance. Predictors were sequentially added and removed based on minimizing the AIC.

Penalized Elastic Net Regression

Another feature selection approach is to apply a shrinkage method, also known as regularization or penalized regression. By applying a shrinkage method, all predictors are present in the model, however, a penalty term causes some predictor coefficients to shrink towards zero, or to become zero. Hence, this extension to the OLS model allows for both variable selection and variable regularization. This method is therefore especially useful in cases where the number of predictors is large. The most common shrinkage methods for linear regression are ridge regression, lasso regression, and EN regression.

Ridge regression is an extension to the OLS model in a way that it adds a penalty term λ that determines whether a predictor coefficient will shrink towards zero. Similar to OLS regression, the goal of ridge regression is to minimize the RSS, but in addition, it also aims to shrink coefficients towards zero depending on the penalty term lambda (λ). The penalty term λ is a tuning parameter: if λ equals zero, then the ridge regression is equivalent to the OLS model. A larger λ, however, will result in more coefficients to be shrunken towards zero. To determine the optimal λ, methods such as cross-validation can be performed, hence the model with the lowest cross-validated error will be selected. The downside of ridge regression is that all predictors stay in the model. Consequently, interpretation of the model may be difficult. For the current case where the aim is to determine which method performs best to predict prices, it is ultimately of substantive importance that pricing practitioners are able to interpret the model well so that it can be utilized as price guidance. Consequently, for the current study, the ridge regression may not yield the most interpretable results. Nevertheless, to solve the interpretation problem, the lasso regression aims to set variables to zero depending on the value of the penalty term λ. Hence, interpretation of the model increases substantively as redundant variables are removed (Tibshirani, 1996). A downside of the lasso regression, however, is that the results do not necessarily define which predictor variable causes the response variable. The reason for this is that in situations with highly correlated predictors, only one predictor variable will be kept in the model (Zou & Hastie, 2003). In addition, the authors note that in cases where a large number of variables are correlated, ridge regression performs better.

Zou and Hastie (2003) have therefore introduced another penalized regression method, namely the EN, which combines the advantages of both ridge regression (e.g. variable shrinkage) and lasso regression (e.g. variable selection). The EN regression aims to solve the following optimization problem (equation 5):

(5)

$$L(\beta_1, \cdots, \beta_p) = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2\right)$$

Where β_j is the coefficient for predictor j (where j = 1, …, p), x_{ij} is the predictor j for observation i, y_i is the dependent variable for observation i. The first part of equation 5, $\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$, minimizes the RSS, hence is equivalent to the OLS model. The second part of equation 5, $\lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + (1-\alpha)\sum_{j=1}^{p}\beta_j^2\right)$, serves as a penalty term and contains two tuning parameters, namely λ and α . First, λ controls how many variables will be shrunken towards zero. Second, α is a tuning parameter between 0 and 1 and determines the extent to which the EN regression will combine the lasso and the ridge regression. For example, is α is equal to 1, the EN regression will perform lasso regression only, while if α is equal to 0, the EN regression will perform ridge regression. In other words, α determines whether or the extent to which the ridge regression and lasso regression will be combined (Zou & Hastie, 2003).

To find the optimal combination of λ and α, for the current study a tuning grid was created using the `caret` package in R. The tune length was set to 25, meaning that in total 25 different combinations of λ and α were computed. The model that minimized the cross-validated error was chosen as the final model. Moreover, the EN regression model was trained using 10-fold repeated cross-validation as a resampling method, where the number of repeats was set to 5. The RMSE was used to select the optimal model, using the smallest value. Moreover, none of the variables were standardized beforehand, because all variables were measured on the same scale.

Random Forest Regression

RF regression is a machine learning technique that falls under the decision tree methods (Breiman, 2001). The RF model is a non-parametric approach, as it does not assume that the independent variable is linearly related to the dependent variables. As a result, prediction tends to be more accurate due to the higher flexibility of the model (James et al., 2013). The RF algorithm sequentially builds multiple fully grown decision trees, predicts unseen data on each tree, after which it aggregates all results to fit the final model by taking the average RMSE. By employing such an approach, the estimation accuracy will be improved compared to a single decision tree, at the cost of interpretability (James et al., 2013). A resampling method such as cross-validation can be used to build the trees. With the RF algorithm, each time a tree is built, a random set of selected predictors will be used, and the predictor that performs best will be used to split the node (Liaw and Matthew, 2002). The parameter that determines how many variables are taken into consideration at each split can be tuned to find the optimal value. A major advantage of this approach is that decorrelated trees are obtained, as no strong predictor will overpower the results, and hence the variance of the model decreases (James et al., 2013).

For the present study, the RF model was trained using the `caret` package in `R`. The model was trained using 10-fold cross-validation, and in total 500 trees were built. To find the optimal number of randomly selected predictors that were used at each split, i.e. tuning parameter `mtry`, the tunelength was set to 15, meaning that 15 random searches were performed to find the optimal value for `mtry` that minimized the RMSE. The final model was chosen based on the `mtry` that minimized the RMSE.

K-Nearest Neighbors

Regression KNN is a non-parametric approach, as it does not assume that the independent variable is linearly related to the dependent variables. As a result, prediction tends to be more accurate due to the higher flexibility of the model (James et al., 2013). The KNN regression algorithm makes predictions based on the *K* closest neighbors. More specifically, a (6) new sample is predicted based on the mean of the *K* closest surrounding observations from the training data (Kuhn & Johnson, 2013). Mathematically, this can be shown by equation 6:

(6)

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

Where, $\hat{f}(x_0)$ is the predicted value for observation x_0, *K* is the number of neighbors, and $\mathcal{N}_0$ defines the *K* closest observations from the training data. For the current study, the distance between samples was defined by the Euclidean distance, shown by equation 7.

$$\left(\sum_{j=1}^{P} \left(x_{aj} - x_{bj} \right)^2 \right)^{\frac{1}{2}} \qquad (7)$$

Where x_{aj} and x_{bj} are two observations for predictor *j*.

Next, the `caret` package in `R` was used to train the model and to determine the optimal value for *K*, using 10-fold cross-validation as a resampling method. The tuneLength was set to 15, meaning that 15 random values of *K* were computed. The optimal value of *K* was selected that corresponded to the model with the lowest RMSE.

Neural Networks

The regression NN algorithm is also a non-parametric predictive modeling technique. A NN is built out of one or multiple hidden units that represent predictor variables that have been transformed by a sigmoidal function. The relationship between the hidden units and the predictor variables is non-linear, however, the hidden units and the outcome variable are linearly related (Kuhn & Johnson, 2013). The number of hidden units is a tuning parameter that can be optimized by the user. In addition to the number of hidden units, another tuning parameter that needs to be defined is the weight decay. The weight decay parameter employs a similar function as the penalty term that was introduced in penalized regression to penalize certain predictors. The aim of incorporating the weight decay parameter in the model is to reduce the potential of overfitting of the model (Kuhn & Johnson, 2013).

For the present study, a single hidden layer NN model was trained using the `caret` package in `R`. To train the model, 10-fold cross-validation was used as resampling method, and the final model was determined using the RMSE. For tuning the parameter representing the number of different units, a range of values from one to ten were tested. Besides the number of hidden units, the values 0.01, 0.1, and 0.5 were evaluated for the weight decay parameter, as suggested by Kuhn and Johnson (2013).

Results

This section will elaborate on the results obtained from the following six models: MPL, SW, EN, RF, KNN, and NN. The following two points will be addressed in this section. First, an overview of the results will be given from which the key findings will be addressed. Second, an example of a price guidance will be provided given the obtained results. The results of the tuning parameters for the EN, RF, KNN and NN can be found in the Tables 3-10 and Figures 4-9 in the Appendix.

Table 2 summarizes the performance of all six models computed on the three data samples. From Table 2, it can be observed that the performance on the test and train set are similar to each other, indicating that no overfitting is present. Overall, the results provide several key insights when comparing the model performances across the different data samples with each other. First, in terms of the test R^2, across all three data samples it can be observed that all models perform superior on subset 1, which removed categories that contained less than 5% of the total number of observations. These results suggest that the number of observations per category influence the prediction estimation on the studied models. In other words, the obtained results imply that having categories that contain more than 5% of the total amount of observations lead to better prediction of the contribution margin for the studied models.

A further notable finding is that across all data samples, the random forest regression outperforms the other models when looking at both the RMSE and R^2 as evaluation metrics. More specifically, the results reveal that the random forest explains approximately 42%, 63%, and 53% of the variation on the entire data, subset 1, and subset 2, respectively, while the other models appeared to explain substantially less variation in the contribution margin. Besides, as demonstrated in Table 2, the test error for the RF model sequentially resulted to be the lowest when compared to the model performance of the other methods computed within the same data sample.

Moreover, another remarkable finding is that the MLR model and its extensions, namely SW and EN, all perform similarly well across the data samples. Given the comparable performance in predicting the contribution margin, the results demonstrate that neither variable selection, nor penalization improves model fit for the present study.

Finally, a striking finding emerged for the NN model performed on the entire data. For this specific case, in terms of RMSE and R^2, the NN model performs noticeably worse compared to the prediction performance of the other models. The R^2 value for both the test and the train set appeared to be extremely low. Additionally, the RMSE is found to be relatively high. From this result it becomes clear that the NN estimated in this study was not able to predict the contribution margin accurately.

Table 2

Model Performance Summary

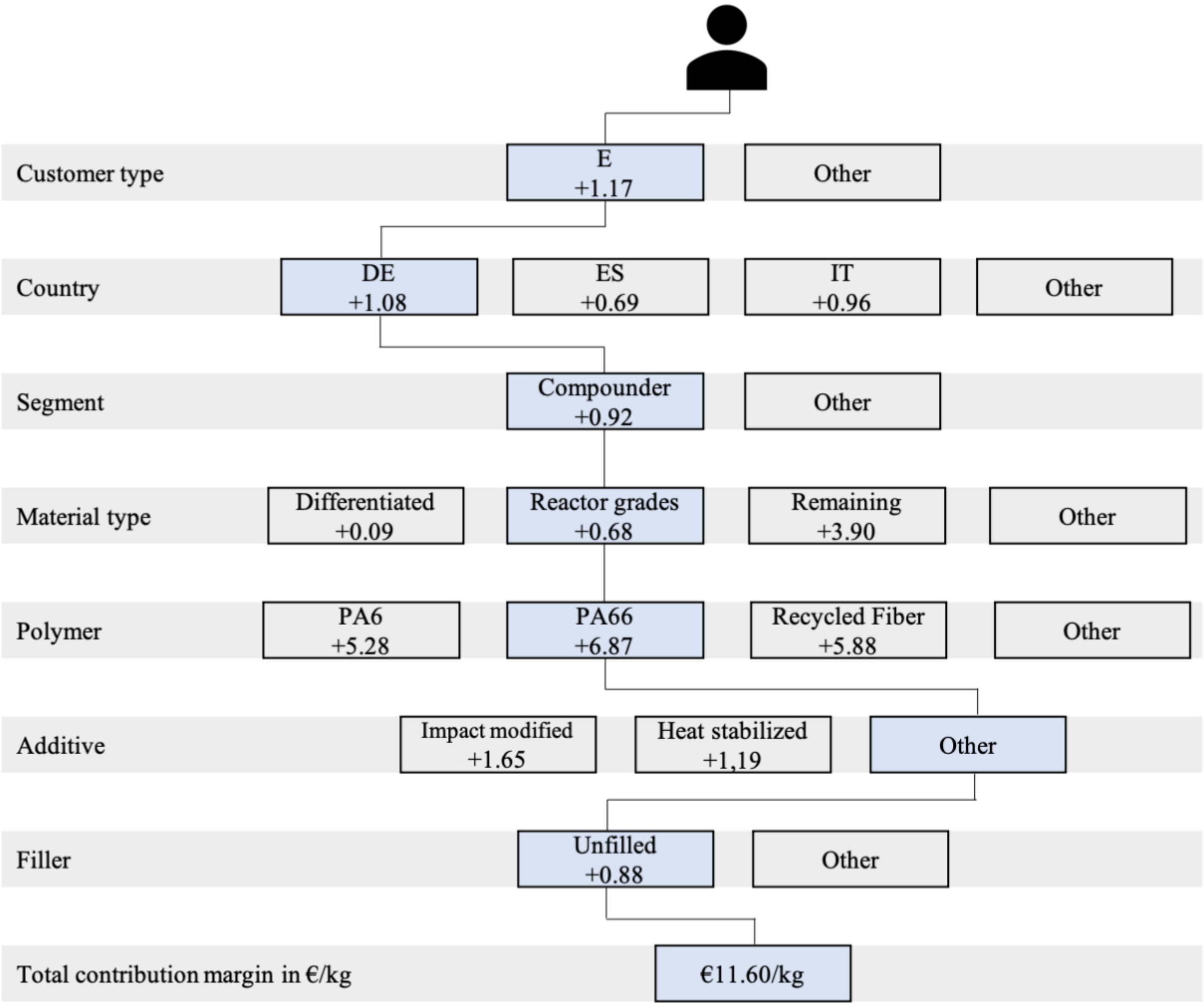
Model	Performance Metric			
	Test		Train	
	RMSE	R^2	RMSE	R^2
Entire data set				
MLR	.604	.363	.613	.402
SW	.604	.362	.613	.402
EN	.606	.354	.643	.344
RF	.578	.422	.585	.458
KNN	.627	.345	.603	.431

NN	.914	.002	.945	.012
Subset 1				
MLR	.350	.524	.373	.493
SW	.349	.527	.493	.373
EN	.493	.525	.488	.385
RF	.311	.626	.456	.464
KNN	.323	.596	.447	.478
NN	.336	.561	.473	.423
Subset 2				
MLR	.211	.377	.207	.389
SW	.212	.379	.207	.387
EN	.210	.382	.210	.371
RF	.183	.528	.187	.500
KNN	.192	.483	.195	.462
NN	.200	.438	.208	.378

Besides looking at statistical significance, from a practical point of view it is also critical to take the interpretability of the models into consideration. For MLR, SW, and EN the coefficients are used for interpretation of the model. For RF, KNN, and NN, however, model agnostic methods are necessary to make the ML model interpretable. Once the results are interpretable, a price guidance can be obtained that help marketeers in setting and getting the right price. For the purpose of this study, an example of a potential price guidance is given in Figure 3 shown below. Figure 3 represents a price guidance with an estimation of the contribution margin in euros per kilograms as output for a given order *i*, that is based on the SW model performed on subset 1. To clarify even further, from Figure 1 it can be seen that if the customer is German, Spanish, or Italian, the contribution margin, hence the price, will significantly increase. In the case shown below, order *i* is predicted to obtain a contribution margin of approximately €11.60/kg, keeping everything else constant.

Figure 3

Example of a Contribution Margin Guidance for Transaction i



Discussion

An extensive review of the current academic literature has shown that statistical models are becoming increasingly popular in today's digital era to optimize pricing decision-making within the marketing domain. The present study has shown that the body of academic literature on data science methods used in pricing is very limited. The majority of pricing literature focuses primarily on pricing strategies. Nevertheless, it became clear that up until now, despite the awareness for the need of more advanced pricing models, scholars have not yet addressed which ML technique would be most suitable for price prediction. Therefore, the present study aimed to investigate the following research question:

"*Which Data Science Technique is Most Effective for Optimal Price Prediction for B2B Businesses?*".

This study aimed to answer the research question stated above by investigating the following three sub-questions:

(1) "What B2B pricing data science methods exist?"

(2) "What data science techniques are relevant for price prediction?"

(3) "How do the methods perform, and which method has the highest performance in estimating the price?"

Based on the literature review, it can be concluded that there are no particular data science methods that are commonly used in price prediction in the B2B sector. This study assessed the performance of six supervised machine learning models, MLR, SW, EN, RF, KNN, and NN, on three data samples to gain a better understanding of why certain models performed the way they did. The remaining of this section will first elaborate on the main findings. Next, the managerial implications of the study will be addressed. After that, the limitations of the study will be mentioned and a suggestion for potential future research will be given.

To give an answer to the research question stated above, from the results it can be concluded that the random forest model was best at predicting the contribution margin. Given the finding that the RF model leads to statistically better results it can be said that the present study obtained similar results when compared to academic literature. Namely, a series of studies obtain similar results when comparing MLR with RF regression in predictive modeling (see for example, Quedrago et al., (2019) and Yuchi et al., (2019)). Besides, since the random forest algorithm makes predictions based on building multiple trees, it is known to reduce the variance substantially (James et al., 2013). Moreover, since the random forest model is a non-parametric approach, it is more flexible in its prediction estimation and therefore tends to obtain higher estimation accuracy with less bias (James et al., 2013).

Furthermore, the results of the present study demonstrated that overall, all models were better at estimating the contribution margin if all categories contained at least more than 5% of the total number of observations. This finding may arise due to the fact that underrepresented categories make it harder for machine learning methods to fit a linear relationship, as parametric methods are sensitive to outliers, hence causing difficulties for the model to find the true relationship. Nevertheless, it was striking that the linear regression models did not perform better on subset 2, which contained only observations that were present within the interquartile range. It was expected that the linear regression models would perform better on subset 2, because for this subset the range of predictor values was much smaller. Hence, one would therefore expect that it would be easier to fit a straight line as outliers were removed. Further research is required to gain a better understanding of why the linear regression models performed the way they did.

Last but not least, the NN resulted to perform extremely poor on the entire data compared to the other models. This could possibly be explained by the fact that the NN model used in this study was not flexible enough to make accurate predictions. For the purpose of this study it was chosen to build a NN model with one hidden layer. Consequently, this may have somewhat biased the results, especially for the NN conducted on the entire data.

Implications

Research in this relatively new area within the marketing domain is especially relevant for pricing practitioners or marketeers who do not yet set and get the right prices. Since the body of academic literature on how to set and get the right prices is relatively scarce, this study can be considered as a starting point for pricing academics to build further on this research, and for businesses to re-evaluate their pricing strategies. Once marketeers gain a better understanding of which data science technique is most accurate in predicting the contribution margin or price, marketeers can subsequently create a price guidance based on the model predictions, given that historical sales data is available. Consequently, pricing practitioners will then be able to charge optimal prices for their customers. More importantly, managers will be able to identify potential price drivers and segments based on product and customer characteristics. Potential price dirvers and segments can be identified by looking at which variables are the key drivers for the price. For example, the results demonstrated that for the studied business, German, Spanish, and Italian customers could be charged significantly more compared to other customers. As a result, managers can create winning deals by setting and getting the right price based on customer and or product segmentation, and thus profit maximization is achieved with higher certainty. In other words, prices can be set more accurately based on customer's WTP.

Limitations

The present study investigated what the effect was of six data science methods on historical sales data from a leading chemical Belgium company operating in the B2B sector. A limitation of this is that the results may not be generalizable businesses operating in different industries or in the B2C sector. Therefore, it cannot be stated with high certainty that the results are valid for all businesses. In addition, since the analysis is based on historical sales data, the study makes the strong assumption that the historical sales data is representable of future prices or contribution margins. This means that external market conditions, such as consumer behavior, are not controlled for, and thus the results may contain an estimation error.

Future research

Investigating whether the price guidance would lead to profit maximization is of substantial importance from a managerial point of view. Hence, a suggestion for future research would be to implement the price guidance to examine whether profit increases. In addition, the studied models could be investigated on different historical sales data, which would also provide more insights on whether the results are also applicable to other businesses operating in different industries. Moreover, further research is required to gain a better understanding of why certain models behave the way they do. Specifically, for the non-parametric methods, model agnostics is necessary to make the results interpretable. This is especially of interest for pricing practitioners, who aim to set prices at customer's highest WTP. Lastly, the present study took the contribution margin instead of the price itself as a dependent variable to minimize the influence of fluctuating variable costs. Consequently, the obtained results predict the contribution margin. A final suggestion for future research is to include variables that account for the fluctuating variable costs. In this way, a more accurate price prediction can be done.

Conclusion

The present research aimed to shed light on data science methods in pricing. It did so by comparing different data science methods in contribution margin prediction. It specifically focused parametric methods (MLR, SW, EN) and non-parametric methods (RF, KNN, NN). Given the results, it can be concluded that the RF model obtained the highest performance in predicting the contribution margin when looking at the RMSE and R^2 as performance metrics.

More specifically, the RF performed best on the data subset that excluded small categories. Hence, the results indicate that having enough observations in each category is important for ML methods to make accurate price predictions. Although the RF performed superior compared to the other models, it raises the question whether similar results would apply when price would be used as dependent variable. Therefore, for future research it should be evaluated whether similar results would be obtained with price as dependent variable, considering also other variables that can account for the fluctuations in for example the raw material costs. Altogether, this research aimed to contribute to current academic literature by adding to the small body of literature on pricing analytics by assessing multiple machine learning methods.

References

Baker, W. L., Marn, M. V., & Zawada, C. C. (2010). The price advantage (Vol. 535). John Wiley & Sons.

Borden, N. H. (1964). The concept of the marketing mix. Journal of advertising research, 4(2), 2-7.

Bouter, E. J. (2013). Pricing, the Third Business Skill: Principles of Price Management. FirstPrice.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Buffett, W. (2011). Pricing power most important in business: Buffett. The Economic Times. https://economictimes.indiatimes.com/news/international/pricing-power-most-important-in-business-buffett/articleshow/7525729.cms

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific Model Development Discussions*, *7*(1), 1525-1534.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623.

Chui, M., Kamalnath, V., & McCarthy, B. (2018). An executive's guide to AI. Retrieved April 2, 2021, from https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/an-executives-guide-to-ai

Davenport, T. H. (2006). Competing on analytics. Harvard business review, 84(1), 98.

Ettenson, R., Conrado, E., & Knowles, J. (2013). Rethinking the 4 P's. Harvard Business Review, 91(1), 26-27.

Ghatak, A. (2017). Machine learning with R. Singapore: Springer.

Hinterhuber, A. (2008). Customer value-based pricing strategies: why companies resist. Journal of business strategy.

Hinterhuber, A., & Liozu, S. (2012). Is it time to rethink your pricing strategy. MIT Sloan management review, 53(4), 69-77.

Ingenbleek, P. (2007). Value-informed pricing in its organizational context: literature review, conceptual framework, and directions for future research. Journal of Product & Brand Management.

Ingenbleek, P., Debruyne, M., Frambach, R. T., & Verhallen, T. M. (2003). Successful new product pricing practices: a contingency approach. Marketing letters, 14(4), 289-305.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert systems with applications, 31(1), 101-107.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.

Kumar, V. (2015). Evolution of marketing as a discipline: What has happened and what to look out for. Journal of Marketing, 79(1), 1-9.

Kumar, V., Ramachandran, D., & Kumar, B. (2021). Influence of new-age technologies on marketing: A research agenda. Journal of Business Research, 125, 864-877.

Kumar, V., Shah, D., & Venkatesan, R. (2006). Managing retailer profitability—one customer at a time!. Journal of Retailing, 82(4), 277-294.

Lancioni, R. A. (2005). A strategic approach to industrial product pricing: The pricing plan. Industrial marketing management, 34(2), 177-183.

Liaw, A. W. (2002). Matthew Classification and Regression by randomForest. R news, 2(3), 18-22.

Liozu, S. M., Hinterhuber, A., Boland, R., & Perelli, S. (2012). The conceptualization of value-based pricing in industrial firms. Journal of Revenue and Pricing Management, 11(1), 12-34.

Martin, G. (2011). The importance of marketing segmentation. American Journal of Business Education (AJBE), 4(6), 15-18.

Miller, K. M., Hofstetter, R., Krohmer, H., & Zhang, Z. J. (2011). How should consumers' willingness to pay be measured? An empirical comparison of state-of-the-art approaches. Journal of Marketing Research, 48(1), 172-184.

PricewaterhouseCoopers. (2021). PwC's Pricing Capability Framework.

Salesforce. (2020). Pricing Guidance. https://help.salesforce.com/articleView?id=sf.cpq_pricing_guidance_intro.htm&type=5

Schindler, R. M., (2011). Pricing strategies: a marketing approach. SAGE Publications.

Shah, D., & Murthi, B. P. S. (2021). Marketing in a data-driven digital world: Implications for the role and scope of marketing. Journal of Business Research, 125, 772-779.

Shah, D., & Shay, E. (2019). How and why artificial intelligence, mixed reality and blockchain technologies will change marketing we know today. Handbook of Advances in Marketing in an Era of Disruptions, 377-390.

Shah, D., Rust, R. T., Parasuraman, A., Staelin, R., & Day, G. S. (2006). The path to customer centricity. Journal of service research, 9(2), 113-124.

Sheth, J. N., & Uslay, C. (2007). Implications of the revised definition of marketing: from exchange to value creation. Journal of Public Policy & Marketing, 26(2), 302-307.

Sondergaard, P. (2011). Is Data Really the New Oil in the 21st Century?. Retrieved March 30, 2021, from https://towardsdatascience.com/is-data-really-the-new-oil-in-the-21st-century-17d014811b88

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. Journal of Marketing, 80(6), 97-121.

Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., ... & Allen, R. W. (2019). Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. *Environmental pollution*, *245*, 746-753.

Zou, H., & Hastie, T. (2003). Regression shrinkage and selection via the elastic net, with applications to microarrays. JR Stat Soc Ser B, 67, 301-20.

Appendix

Table 1a

Variable Definitions

Variable	Definition
Material Type	Type of material used to manufacture the product
Polymer	Type of polymer used to manufacture the product
Additive	Specificity of the type of material used to manufacture the product
Filler	Type of filler used to manufacture the product
Color	Color of the product
Segment	Segment the customer is operating in
Customer	Customer classification
Country	Country where the order was shipped to
Customer volume	Total annual customer volume

Table 1b

Customer Definitions

Customer Type	Definition	Broader Description
A	Global Strategic Customer	High Potential and Global Accounts
B	Local Valued Customer	Willing to pay for quality and service
C	Local Price Buyer	Mainly interested in price
D	Non-Strategic Customer	Typical small volumes
E	Strategic Distributor	Key distributor willing to grow with
F	Non-Strategic Distributor	Distributor with limited potential
G	Compounder	EP competitor

T	Trader	Export markets
Z	Other	None of the other classifications

Table 1c

Segment Definitions

Segment	Description
Auto-Mix	Automotive industry
E&E	Electric & Electronics
ICG	Industrial & Consumer Goods

Table 2

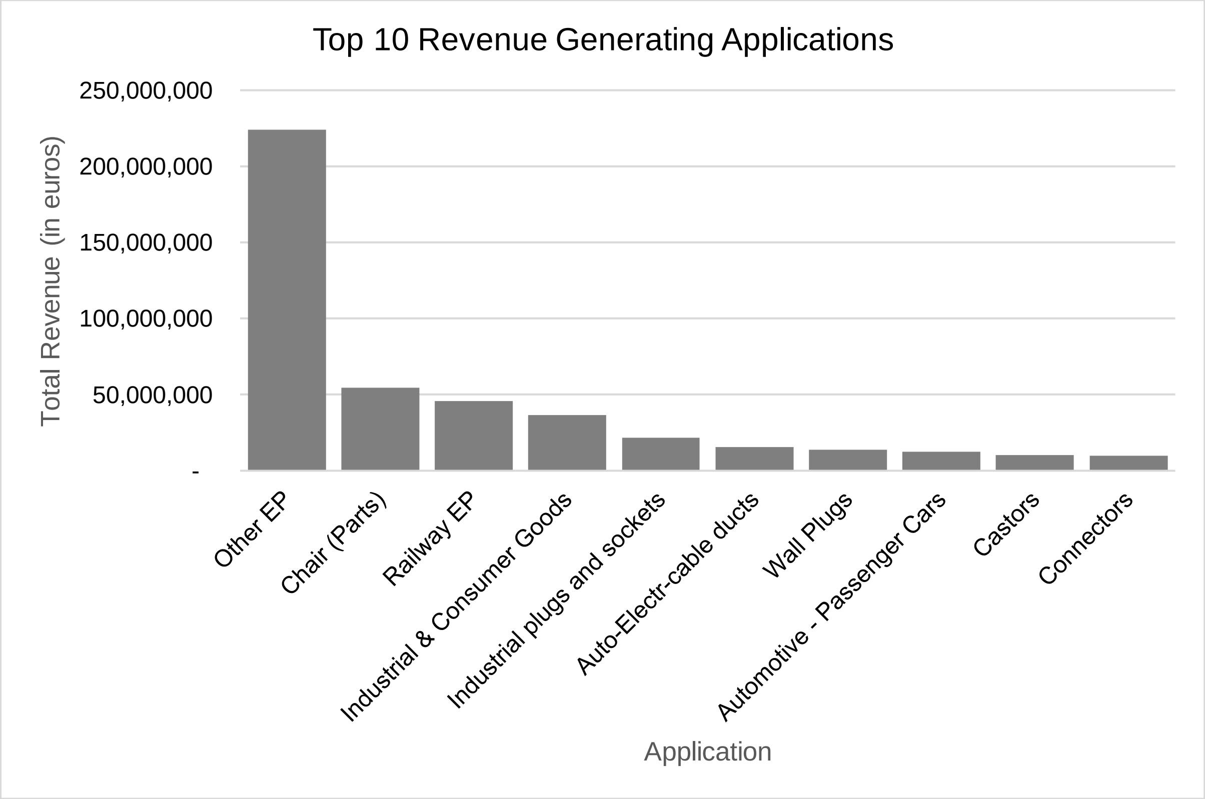
Descriptive Statistics of the Numerical Variables

	Volume	Total Revenue	Sales Costs	Contribution Margin	Total Customer volume	Price
Min	0	0	0	0	0	0
Median	3925	9475	39.43	2172	3925	9475
Mean	6565	18125	246,11	6897	6565	18125
Max	28075	47850000	3796,88	47810510		

Max	28075	47850000

Figure 1

Bar Chart of the Top Ten Revenue Generating Applications



Note. Based on net revenue.

Figure 2

Distribution of the Dependent Variable Contribution Margin

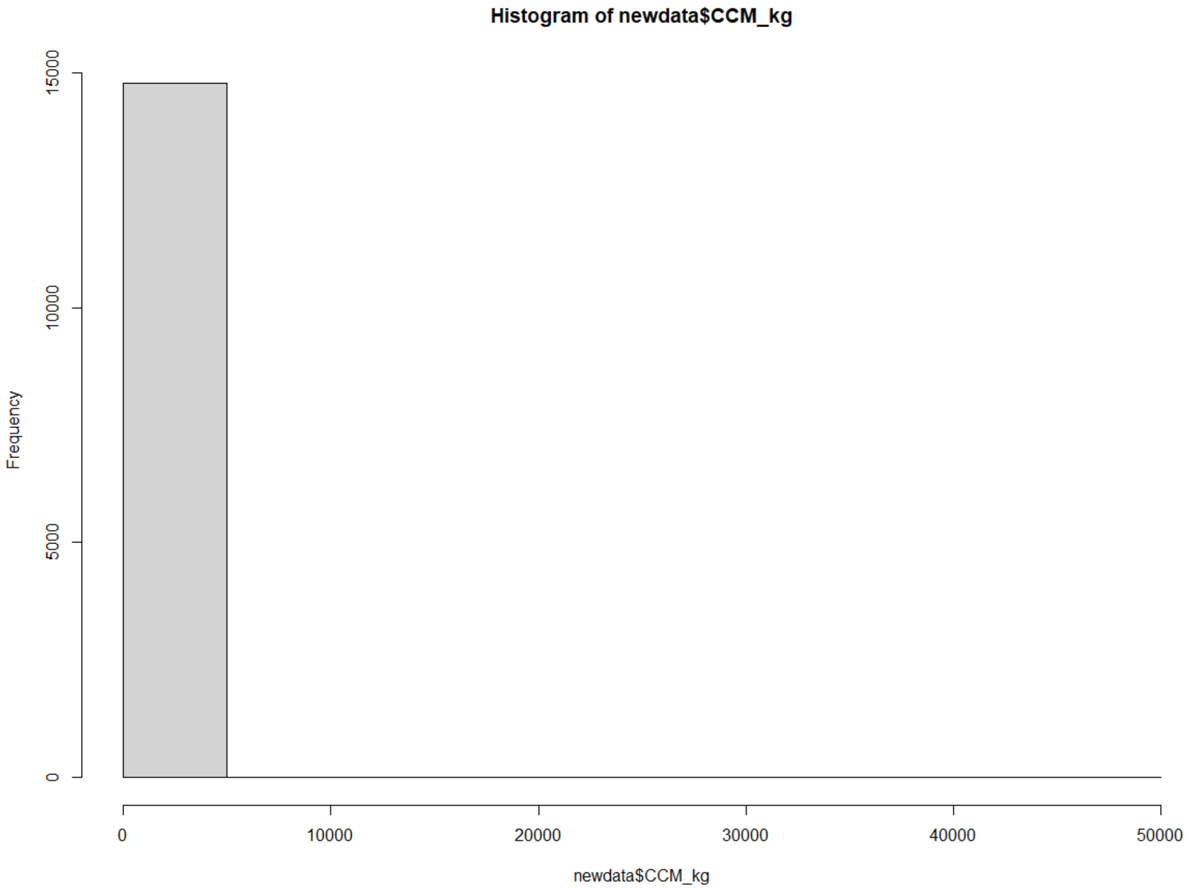


Figure 3

Distribution of the Log Transformed Dependent Variable Contribution Margin

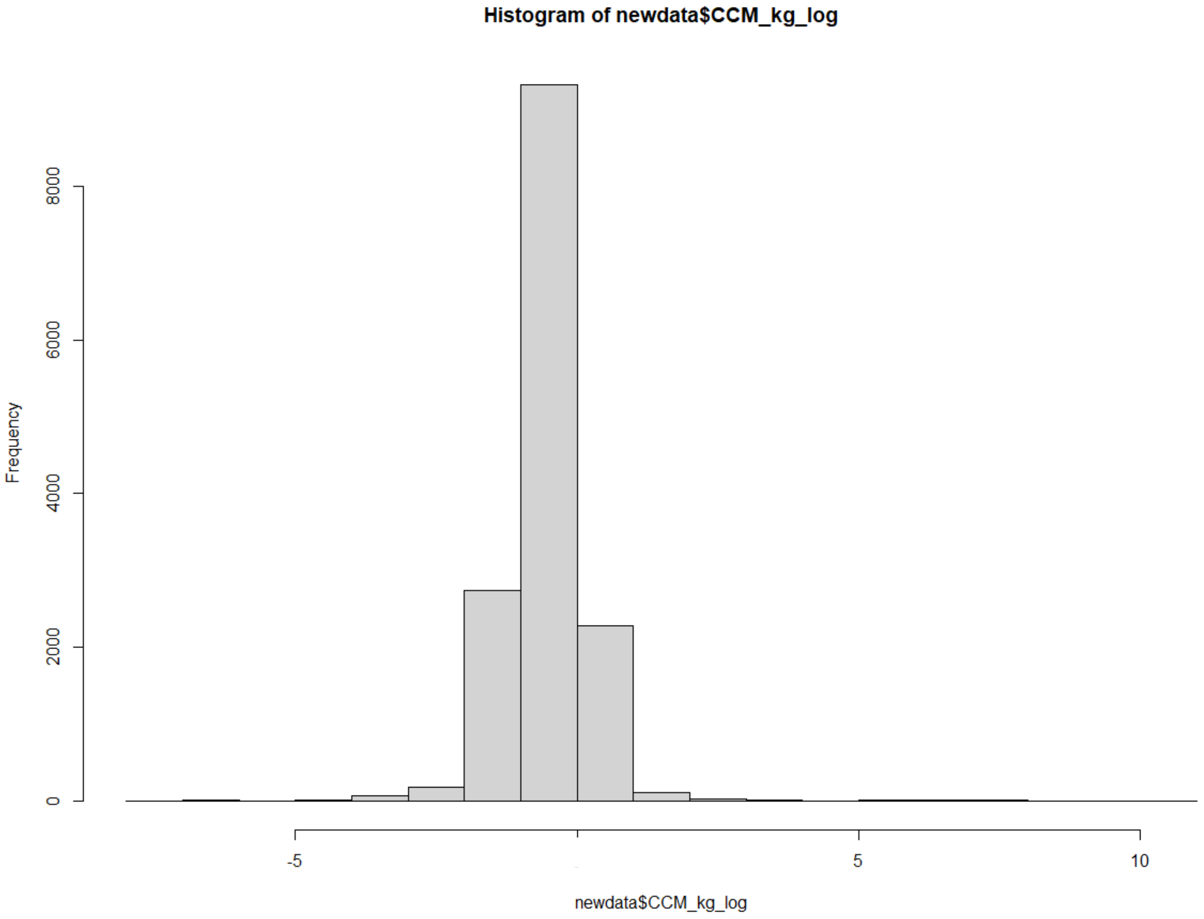


Table 3

Optimal Tuned Parameter Results of Elastic Net on entire Dataset

Alpha	Lambda	RMSE	R-squared	MAE
.906	.0010	.606	.354	.329

Table 4

Optimal Tuned Parameter Results of Random Forest on entire Dataset

mtry	RMSE	R-squared	MAE
40	.578	.422	.272

Table 5

Optimal Tuned Parameter Results of KNN on entire Dataset

k	RMSE	R-squared	MAE
5	.627	.345	.291

Table 6

Optimal Tuned Parameter Results of NN on entire Dataset

Size	Decay	RMSE	R-squared	MAE
7	.1	.914	.002	.679

Figure 4

Parameter Tuning Results of K-Nearest Neighbors on Entire Dataset

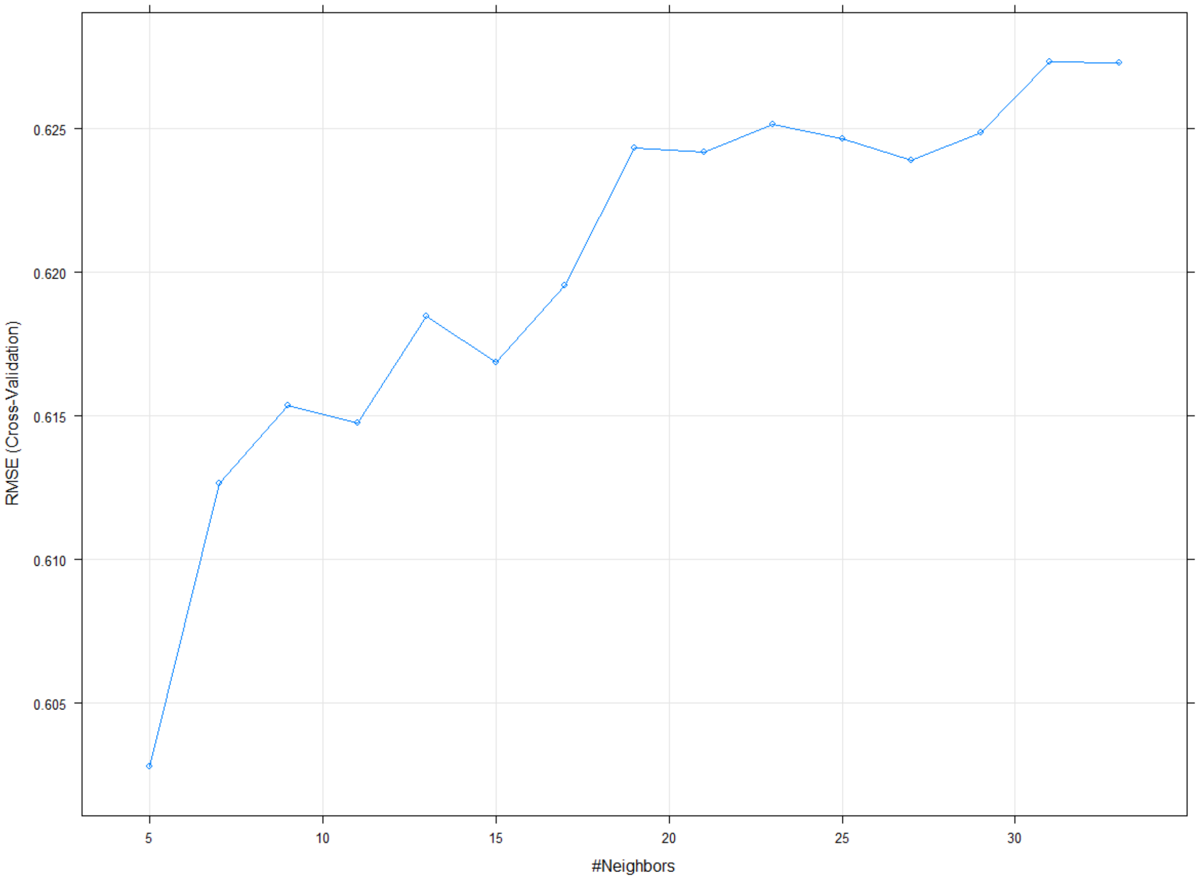


Figure 5

Parameter Tuning Results of Neural Network on Entire Dataset

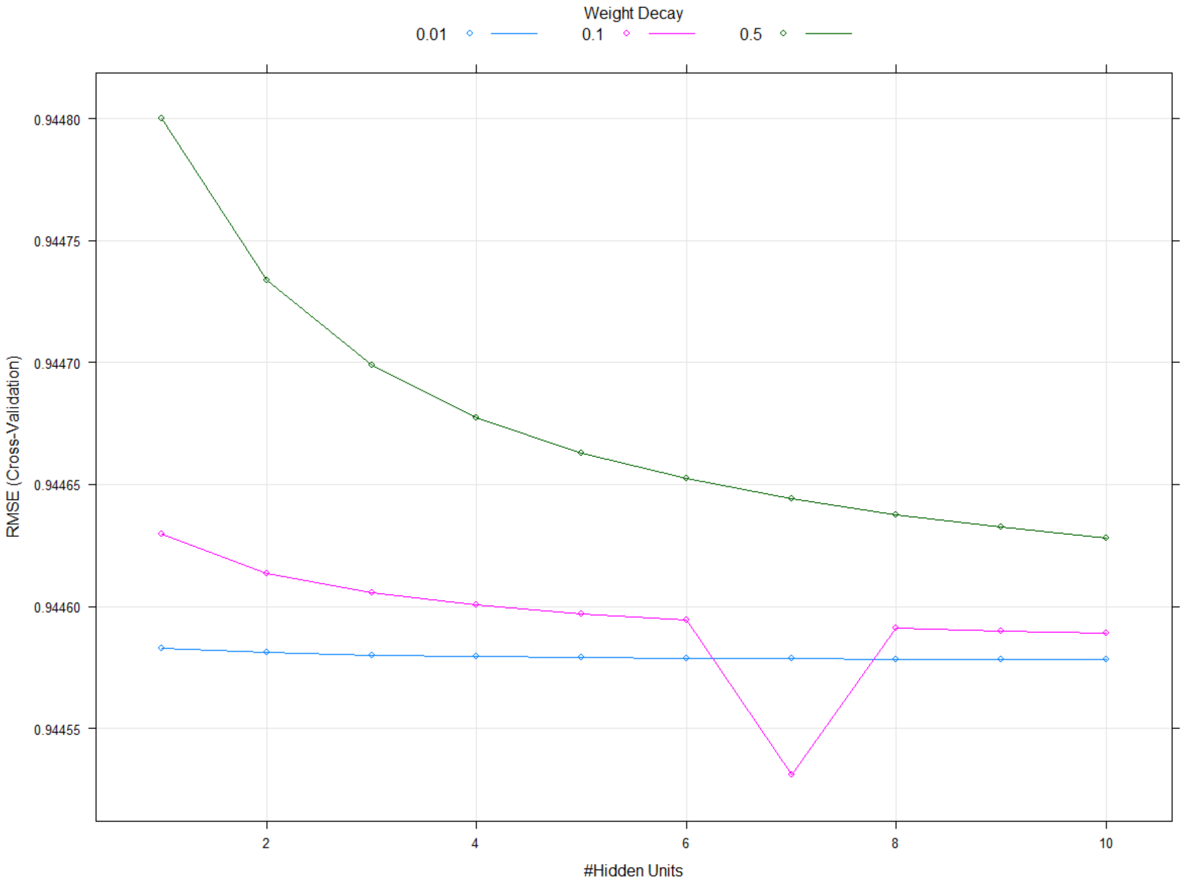


Table 7

Optimal Tuned Parameter Results of Elastic Net on Subset 1

Alpha	Lambda	RMSE	R-squared	MAE
.255	.004	.493	.525	.283

Table 8

Optimal Tuned Parameter Results of Random Forest on Subset 1

mtry	RMSE	R-squared	MAE
15	.311	.626	.249

Table 9

Optimal Tuned Parameter Results of KNN on Subset 1

k	RMSE	R-squared	MAE
5	.323	.596	.256

Table 10

Optimal Tuned Parameter Results of NN on Subset 1

Size	Decay	RMSE	R-squared	MAE
9	.5	.336	.561	.274

Figure 6

Parameter Tuning Results of K-Nearest Neighbors on Subset 1

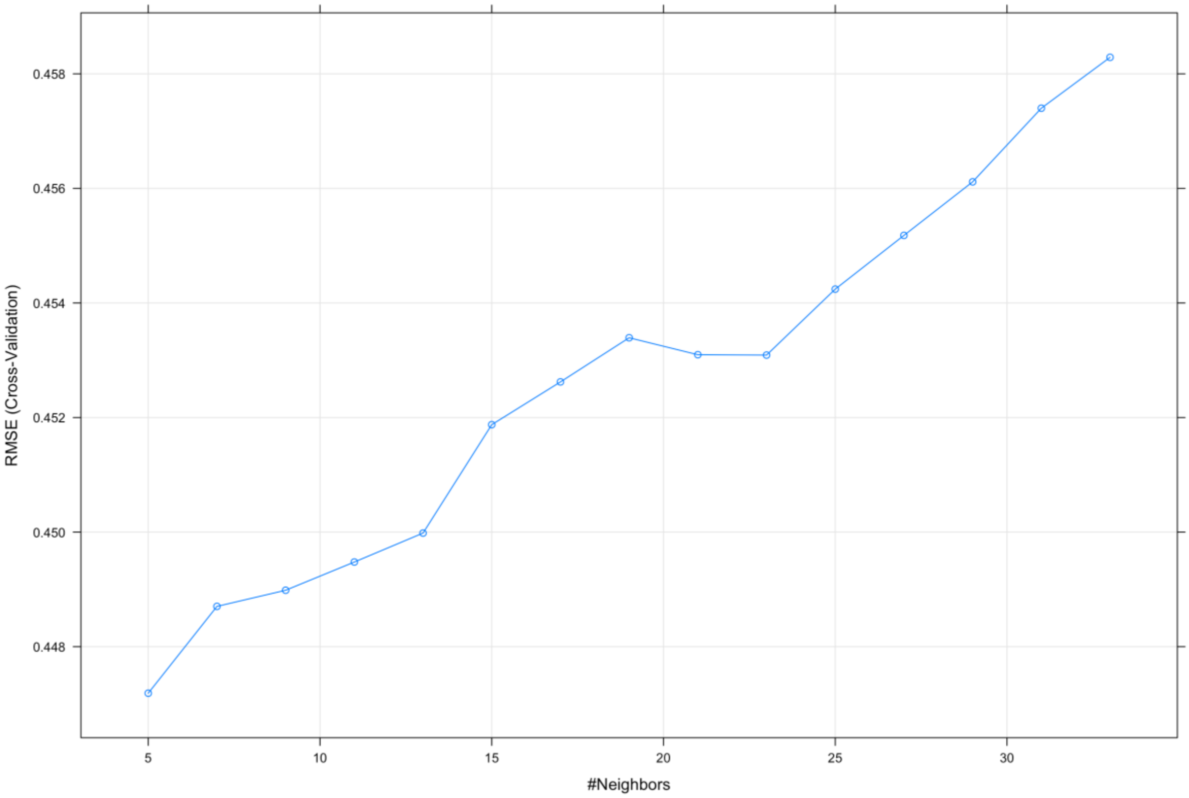


Figure 7

Parameter Tuning Results of Neural Network on Subset 1

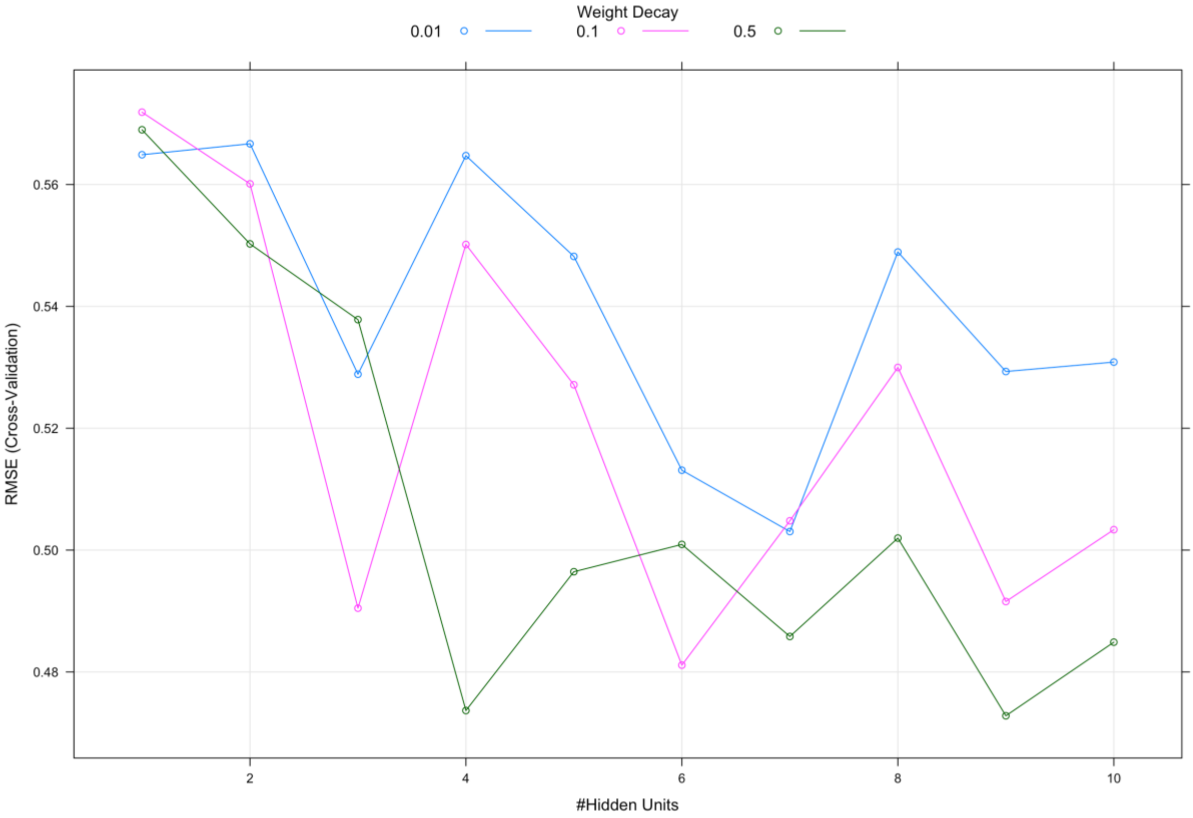


Table 7

Optimal Tuned Parameter Results of Elastic Net on Subset 2

Alpha	Lambda	RMSE	R-squared	MAE
.220	.003	.210	.382	.167

Table 8

Optimal Tuned Parameter Results of Random Forest on Subset 2

mtry	RMSE	R-squared	MAE
40	.183	.528	.142

Table 9

Optimal Tuned Parameter Results of KNN on Subset 2

k	RMSE	R-squared	MAE
5	.192	.483	.149

Table 10

Optimal Tuned Parameter Results of NN on Subset 2

Size	Decay	RMSE	R-squared	MAE
10	.5	.200	.438	.165

Figure 8

Parameter Tuning Results of K-Nearest Neighbors on Subset 2

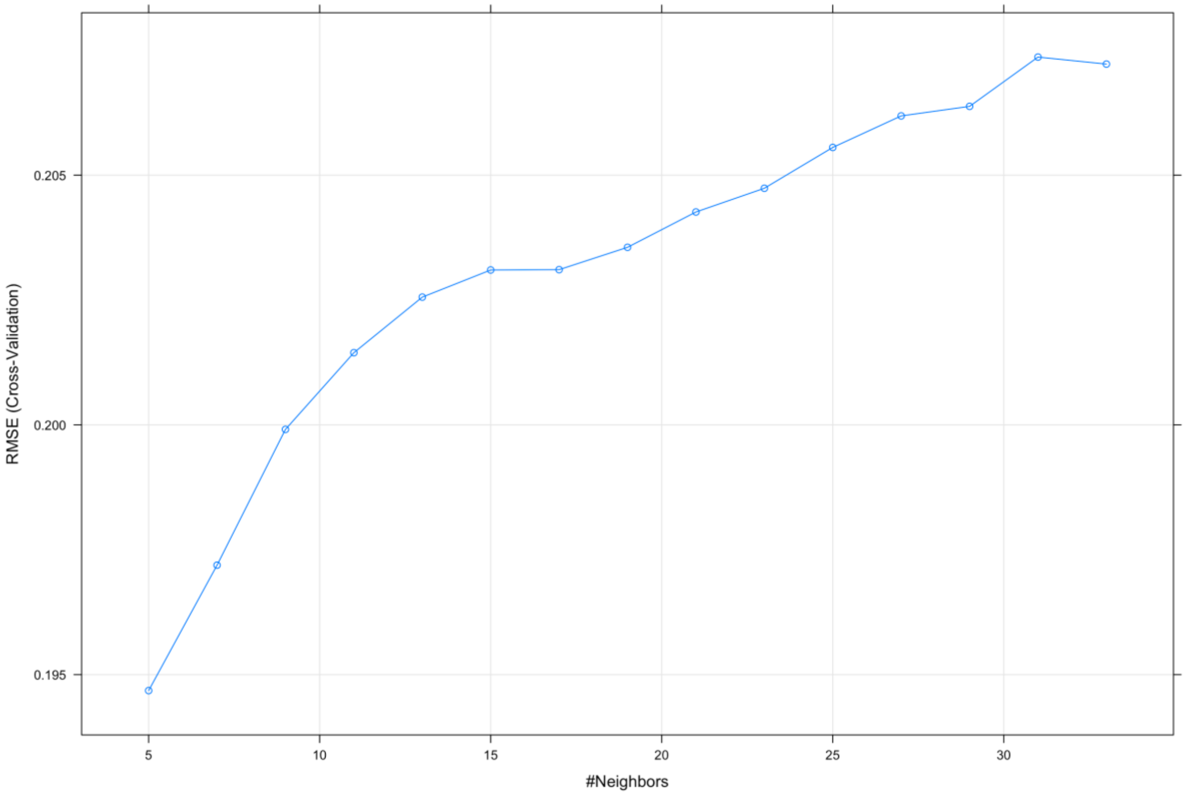
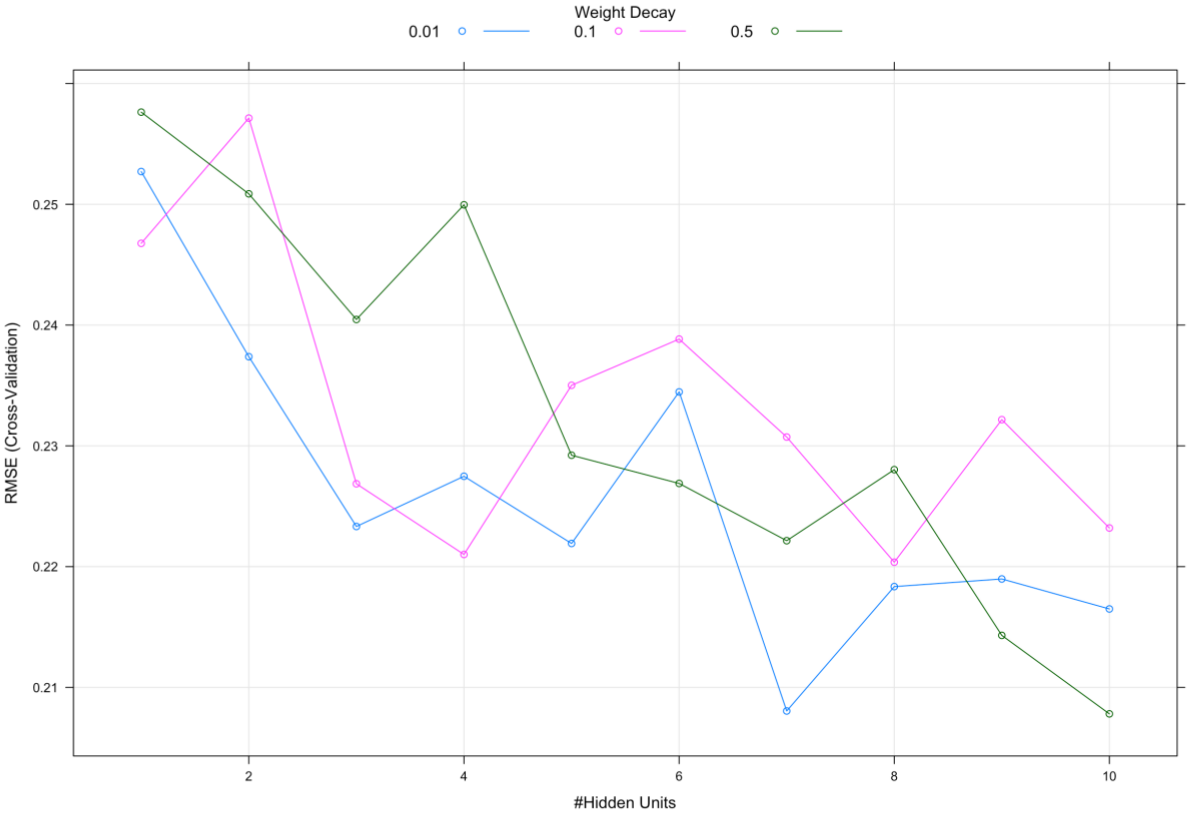


Figure 9

Parameter Tuning Results of Neural Network on Subset 2


Weight Decay
RMSE (Cross-Validation)
#Hidden Units