Norwegian School of Economics

Bergen, Spring 2021

Level Up Your Sneaker Game

Applying machine learning techniques to support data-driven investment decisions in the sneaker resale market

Sandrina Kenny, Asli Cetin

Supervisor: Håkon Otneim

Master thesis, Msc in Economics and Business Administration,

Major in Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

This thesis was written as a part of the Master of Science in Economics and Business Administration at NHH. Please note that neither the institution nor the examiners are responsible − through the approval of this thesis − for the theories and methods used, or results and conclusions drawn in this work.

Abstract

Sneaker resale has become a worldwide phenomenon. The resale market is growing, expected to potentially reach up to $30 bn by 2030. More and more people want to take part in making fortunes out of shifting high valued Nike SB sneakers, rare Air Jordans or eccentric Yeezys. Notably, traditional customer roles are changing: consumers are no longer only buying sneakers for wearing them themselves, but are also engaging in resale activities. Additionally, new market participants are entering the game with the sole aim of making profits as large as possible from buying and then reselling brand new shoes.

The purpose of this thesis is to provide insight into how machine learning methods can support data-driven investment decisions in the sneaker resale market. Two different reseller personas will be introduced, together with a description of scenarios and questions these might encounter.

Using data from StockX.com, the leading marketplace for sneaker resale, various machine learning techniques will be applied to arrive at founded investment decision for these two personas. To meet the needs of the different personas, this thesis makes use of both simpler methods, such as linear and logistic regression as well as KNN and regression trees, and more complex methods such as Random Forest and XGBoost models.

The authors chose a practical approach with the analysis of different scenarios, aiming to allow sneakerheads, who engage in and are hence interested in information on resale markets, to profit from the insights.

The research shows that both simple and complex methods can be useful in these decisions, reaching high accuracy values as well as oftentimes good predictions. It also shows that the sneaker resale price is influenced by a myriad of factors, and that especially celebrity collaborations seem to have high influence on resale value of sneakers.

Acknowledgments

The authors wrote this thesis during fall of 2020 and spring of 2021 at the Norwegian School of Economics (NHH). It is part of their MSc in Economics and Business Administration, with specialization in Business Analytics.

Despite the challenging times and lack of possibility to work together physically, the authors are thankful to have been able to pursue this project. The authors aimed to create a both exciting and interesting thesis on a topic that, to their knowledge, has not been discussed in this extent before.

The authors would firstly like to thank their supervisor, Associate Professor Håkon Otneim, very much for his guidance and support during this project and for enabling them to explore this potentially unconventional thesis topic.

Furthermore, the authors would like to thank their families and friends for their support, advice and encouragement along the way. Very special thanks are extended to Farzan and Jason – thank you for keeping our moods up during this time!

Contents

List of Tables

List of Figures

List of Abbreviations

ACC	Accuracy rate
ANN	Artificial neural networks
AUC	Area under the receiver operating characteristic curve
CV	Cross validation
ERR	Error rate
FN	False negatives
FNR	False negative rate
FP	False positives
FPR	False positive rate
KNN	K-nearest neighbors
MSE	Mean squared error
NN	Neural networks
RF	Random Forest
RMSE	Root mean squared error
ROC	Receiver operating characteristic
RQ	Research question
RSE	Residual standard error
RSS	Residual sum of squares
sku	Stock-keeping unit
TN	True negatives

TNR	True negative rate
TP	True positives
TPR	True positive rates
TSS	Total sum of squares
XGBoost	Extreme gradient boosting

Glossary

Back-dooring

Back-dooring refers to the practice of resellers purchasing large amounts of limited sneakers via unofficial trade routes and personal relations. This often takes places before the official release date of a sneaker.

Deadstock

Sneakers that are authentic, new, unworn, and that come with their original packaging, are referred to as deadstock in the secondary market for sneakers.

GOAT

GOAT is an online marketplace for primarily sneaker resale, which allows sellers to list their sneakers, and buyers to purchase the listed items.

Mixed-Role Reseller

A mixed-role reseller is an individual who initially purchases limited-edition sneakers for their own use, e.g. for their sneaker collection, but aims to resell at a profit or to minimize loss after a certain holding period.

Online raffle system

With an online raffle system, a retailer provides customers the chance to win the right to purchase a sneaker.

Price Premium

Price premium is a measure of the value increase of a sneaker in the secondary market. It measures the total margin, compared to the retail value, a reseller can achieve through the resale of a sneaker.

Professional Reseller

A professional reseller is an individual who purchases limited-edition sneakers with the sole reason to maximize profits through resale on the secondary market.

Resale Price

The resale price is the value of a sneaker in the secondary/resale market.

Reseller

In the context of sneakers, reseller is referred to an individual who engages in investment activities in the sneaker resale market, that is, purchasing and reselling limited-edition sneakers with the intention to generate profits.

Retail Price

The retail price is the initial price of a sneaker, as set by the brand, in the primary market.

Silhouette

Silhouette is the shape/model of a sneaker.

Sneakerhead

Sneakerhead is referred to an individual who is passionate about collecting sneakers and who assigns a greater value to sneakers than the wider population.

Sneaker Resale Market

The sneaker resale market is a secondary market for limited-edition sneakers.

StockX

StockX is the leading online marketplace for sneaker resale. The platform resembles a stock exchange platform, with sellers setting their prices and buyers bidding for the sneakers.

1. Introduction

The functionality and usage of sneakers has been changing, and sneakers are no longer worn for athletic performance purposes only. Today, sneakers are established as a fashionable item and seen as a status symbol. With that, a new consumer group for sneakers emerged, referred to as *"sneakerheads"*. According to Choi (2017), sneakerheads are people "who assign much more value to sneakers than the general population, and collect them with great passion" (p.1).

These sneakers, which attract the attention of sneakerheads, are typically limited in numbers and sell out within mere seconds after release. Hence, collectors typically have little chance to purchase sneakers from the primary market, and turn to resale marketplaces, where the prices oftentimes increase significantly above the initial retail prices. It can be assumed that the current market phenomena are results of the development of sneakers as a symbol of status throughout history. In order to analyze the price drivers of sneakers in the resale market, it is hence crucial to gain an understanding of the development over time of the sneaker culture as well as of major brands and collaborations. Therefore, the following section aims to summarize the evolution of this exceptional footwear. This is followed by a section that gives an overview of today's online sneaker resale market, and thereafter, a section that presents the research questions addressed in this thesis.

1.1 History of the Sneaker Culture – From Early Beginnings to Today's Hype

Early stages – Sneakers as practical shoes for leisure and sports

The first shoes used in sports and leisure were mainly intended for reasons of practicality, limiting the damage caused to communal croquet lawns (Denny, 2020). The industrial revolution, allowing for cheaper mass production, facilitated the production of the first rubber sole as well as growth of prosperity and consequently more opportunities for leisure time amongst the public. Yet, the shoes remained mainly functional, supporting the desire of consumers for comfortable and practical shoes (Semmelhack et al., 2015; Denny, 2020).

Following the invention of basketball at the end of the 19th century, rubber soled shoes became the new norm for all court-based sports (Denny, 2020). In 1917, the first Converse All Star sneaker was released, the first basketball shoe offered by the company. Unusual at the time, Converse engaged in both a first collaboration with a basketball coach for promoting the shoe and what can be seen as first celebrity endorsement by hiring basketball player and coach Chuck Taylor (Semmelhack et al., 2015; Denny, 2020). This provided the means for the Converse Chuck Taylor All-Star to become and remain one of the most well-known sneakers in the world (Denny, 2020).

After the stock market crash in 1929, thanks to manufacturing innovations and development as well as cheap imports, sneakers began to become a low-cost, casual fashion item worn also for non-athletic purposes (Semmelhack et al., 2015).

Further athlete collaborations – Sneakers as a way to express individuality

In the 1970s, sneakers were seen as a means of expressing individuality, style, status, and masculinity. This was further nurtured with companies supplying wide ranges of possible models and colors, allowing consumers to find their very own preferred style. The cultural trend in American society of allowing individual expression through brand identification matched the increased interest in sneakers (Semmelhack et al., 2015).

In 1972, Puma signed basketball player Walt Clyde Frazier to endorse their sneakers. This collaboration is often referred to as the first true collaboration between a sneaker brand and an athlete. Frazier's success in his sport was not the only reason for this collaboration, but mainly his "outrageously fashionable off-court persona". Having a persona incorporating the mix of athlete, style and star power endorse a shoe supported the development of sneakers away from mere practicality (Denny, 2020).

Following this, athlete endorsements became popular. Nike even managed to persuade former adidas fan Michael Jordan to sign an endorsement deal with them, offering him an entire brand named after him and various embodiments of sneakers, instead of "only" a single sneaker (Denny, 2020). In the 1984-85 NBA season, Jordan wore the black and white color of the Air Jordan 1, and was rebuked immediately by the NBA as the shoe did not respect the "uniformity of uniforms" rule (Semmelhack et al., 2015). Jordan received a fine for wearing them, which became advertising gold for Nike, as Jordan continued to wear the shoes, flaunting a rebellious act, and standing up for his own individual style. This made the Jordan sneaker one of the most important models for sneakerheads worldwide. Following Denny (2020), Nike's collaboration with Michael Jordan is said to be one of the major collaborations that popularized and impacted the sneaker collecting culture.

Brand and celebrity collaborations – Sneakers as a fashionable item

The more society evolved, the more the sneaker was established as a fashionable item and not solely athletic wear. Casual Fridays, as well as e.g. Michael Jordan stating that an Air Jordan 11 would fit with a suit, increased the everyday usability of sneakers. This led to high fashion brands, such as Prada, entering the sneaker game, either through collaborations or their own designs. One of the key collaborations was the Nike x Supreme collaboration, an extremely limited release which caused chaos when customers and fans realized just how limited this release was. Many camped outside retail stores days before the launch of the objects of desire, and outrage caused by the highly limited supply meant the police had to come. This event raised concerns and discussions on sneaker culture and violence, supported by the media titling the news "sneaker riot" (Tsjeng, 2014; Cochrane, 2014).

Collaborations with celebrities such as Kanye West or Jay-Z grew in popularity, while further increasing hype around the sneaker culture through limited release editions. A collaboration with Kanye West hinted at another peculiar characteristic of the sneaker culture: after having partnered with Nike for several years, Kayne West announced his moving to adidas – just shortly before releasing his final sneaker with Nike, the Air Yeezy 2 "Red October", in 2014. Although retail value of the sneaker was set at $250, the resale value on eBay reached up to $90.300. As Semmelhack et al. (2015) point out, for sneakers, the "real value – to all involved – was the amounts they would fetch at resale" (p. 178-183), giving first indications on the price differences between retail and resale prices as well as the importance of the resale market to the sneaker world overall.

Sneaker culture today

The sneaker development and history are of major importance when analyzing the relevance these shoes hold today. The sneaker was originally seen as merely a utensil to facilitate physical action, however, today's sneaker culture puts emphasis on sneakers as status symbol. The sneaker is now less about the actual usage and more about collecting and owning shoes and taking part in a community**.** Although the sneaker today is still connected to traditional sportswear, it has become more of a fashion statement and is considered a decent choice of footwear for both casual and formal attire. The cultural development has gone through a variety of social shifts, focusing mainly on the perception of individualism and status, and challenges the traditional image and understanding of masculinity (Semmelhack et al., 2015).

In the following, an introduction to resale markets in general and the sneaker resale market in particular is given.

1.2 Online Resale Markets and Sneaker Resale Today

After having started off as casual garage sales, with neighbors selling their old, used and unwanted goods to those passing by, the invention of the internet allowed for resale to become more and more prominent across different product groups (Choi, 2017; Chu & Liao, 2010). Reselling describes the act of selling goods from one customer to another, where the seller was himself a customer at either the retailer or another customer at an earlier point in time (Choi, 2017). Hence, one can see that a shift in traditional roles of seller and consumer occurs in the resale market (Choi, 2017). The resale market has been, and is forecasted to continue, growing in size. As the 2020 Resale Report by ThredUp shows, the secondhand apparel market value worldwide has risen from $11 bn to $28 bn in 2019, while predictions estimate these numbers to increase to over $50 bn by 2023 (ThredUp, 2020).

Just as for the resale market for apparel and other products, the development of the internet has opened up new opportunities for sneaker resale. Before, sneakerheads would meet at conventions, reselling their purchases made at thrift stores, flea markets or off the shelves of their nearest retailer (Dayton, 2020).

Nowadays, despite conventions and off-shelf purchasing still taking place, the majority of sneaker resale is performed over online platforms. Allowing more users to participate lead to a continuous growth of the sneaker resale market, making this one of the most powerful resale markets, estimated to reach $30 bn by 2030. Sneakerheads, those market participants who buy, collect and sell sneakers as a passion, are no longer the sole participants in this market. After classifying sneakers as an emerging alternative asset class, Cowen Equity Research noticed a growing and more diversified investor base (Wade, 2021).

Increased interest by investors chasing favorable risk-reward earnings by engaging in a diversified asset class led to the creation of new online marketplaces primarily for sneaker resale such as StockX and GOAT. StockX, a platform founded in February 2016, works differently to traditional online marketplaces such as eBay, where users log on and resell or buy used sneakers. The platform in fact resembles more a stock exchange platform, with sellers setting their prices and buyers bidding for the sneakers. In order to make sure the sneakers sold on StockX are originals, the company employs anti-counterfeiting services, hence the shoes are not sent directly from seller to buyer but are shipped via professional sneakerheads who verify the quality.

The sneaker resale market allows for large profits to be made. Examples for this are e.g. shoes released as collaborations, such as the Nike Air Yeezy 2 “Red October”, adidas Yeezy Boost 350 “Turtledove” or the Air Jordan 1 x Off-White “Chicago”. All three of these were released at retail prices between $190 to $240, while their average resale value increased to astonishing $1,695 - $6,118 (Steinberg, 2018).

Within the sneaker resale market, it seems like the value of sneakers knows no limit. Price increases up to a couple of thousand dollars are not unusual, which leads to new market participants to enter and participate in the market – one being professional resellers that purchase and resell sneakers for the sole reason of maximizing profits.

1.3 Research Questions, Design and Aim

The increased importance and growth of the sneaker resale market follows different and not always clearly identifiable impulses. Some sneakers, as mentioned before, achieve price premiums way above their original prices, while others are resold at prices below their initial retail prices. Further, new developments in the sneaker resale market resulted in the division of resellers into different profiles. With professional resellers entering as market participants, who invest in sneakers to maximize profits, the traditional sneakerhead, who purchases and sells sneakers as a passion, is no longer the sole reseller in the market. This development creates different reseller profiles with diverse motivations and needs in the sneaker investment decision-making process.

The present analysis is dedicated to finding out how the resale price of sneakers is influenced by its features such as brand, model or color. Special focus will lie on the impact of celebrity and other collaborations on resale price and reseller investment decisions. Further, the analysis aims to address the different reseller profiles by focusing on and addressing their diverse needs.

The following research questions (RQ) will be addressed henceforth.

RQ1: By what measures is the resale price of a sneaker impacted?

RQ2: How can machine learning techniques help sneakerheads in their decision to buy a sneaker with the intention to resale at profit?

> **RQ2a:** Which machine learning techniques provide the most accurate predictions for professional resellers investment decisions?
>
> **RQ2b:** Which machine learning techniques provide simple, interpretable models, and yet result in a decent level of accuracy for mixed-role investment decisions?

RQ3: How do celebrity and other collaborations impact the resale price?

This thesis consists of seven sections. The following section reviews existing literature on resale markets, the impact of supply and demand on price developments, and the application of machine learning in price predictions. Based on the findings, the literature review is concluded by formulating several hypotheses to be evaluated in the course of this thesis. Section three describes the machine learning techniques and the methods of model assessment and validation applied. An overview of the dataset and the data preprocessing steps is given in section four, whereas section five describes the author’s approach in exploring the research questions with special focus on RQ2, which includes the introduction of two reseller personas. This is followed by the presentation of the modelling process and the analysis of the model’s results. Section six includes a discussion of the research. Finally, section seven concludes the findings of this thesis together with an overview of limitations and further research areas.

2. Literature Review

In this section, the existing literature is explored in order to understand how resale prices are influenced and how resale prices higher than retail prices come to place. First, previous findings on resale markets as a whole are presented. This is followed by economic theory on the impact of supply and demand on price settings, further followed by previous research on the impact of these two variables on price development. Next, the findings on prior applications of machine learning techniques in the context of price predictions are summarized. Finally, findings from the literature review are concluded and hypotheses to be tested are derived.

2.1 Resale Markets

Dolbec & Parmentier (2019) argue that resale markets are built on a “network of desires”. Finding the fact that mainstream products appreciate in value in the short time between being bought at a store and then sold online puzzling, they classify this a gain “typically reserved for assets, such as houses and high art” (p. 539). In their research, they define three categories of resale products, which are each related to different motives of purchasing in resale markets. These categories are vintage products, bought out of the desire to differentiate together with nostalgic reasons, used products, which are usually bought out with the desire to make a bargain purchase, and flipped products (Dolbec & Parmentier, 2019).

Dolbec & Parmentier (2019) furthermore mention the fact that customers can have negative as well as positive responses to products that have already been touched, e.g. show that objects that have been used by someone who they are fond of are being perceived as of higher value. Hence, the increase in value of a pre-touched product is derived from a symbolic interaction model, which rests upon the desire of people wanting to “be associated with an item that belongs to someone about whom they have strong, positive feelings” (Argo et al., 2006, p. 82). This theory fails to explain the rise and attractiveness of the resale market for sneakers, as the resellers are usually not associated with the product but act as consumers in the market themselves (Choi, 2017).

Dolbec & Parmentier (2019) further research the change of markets due to desire-valuation, including the transformation of desire into value, the borders between the traditional roles of consumers and resellers becoming more blurred, and the increased consumer-based competition. They conclude that the development of these so-called networks of desire can be used to partly explain the hype around sneaker resale, starting from the desire associated with sneaker models to more dynamic market environments due to the transformation of consumers to resellers.

After highlighting most important research on resale markets, the following chapter emphasizes the impact of supply and demand on the resale value of products, and especially sneakers.

2.2 Supply and Demand in the Sneaker Resale Market

Following economic theory, the relationship between supply and demand of a certain good or service impacts its price (Marshall, 1920; Kramer, 2019; Britannica, T. Editors, 2019).

The theory states an inverse relationship between supply of goods and services and their prices in times of constant demand. While demand remains the same, a decrease in supply will lead to rising prices and lower quantities sold, while an increase in supply will lower prices. The same concept holds vice versa, as a decrease in demand will lower prices while an increased demand will, should supply remain constant, increase prices (Marshall, 1920; Kramer, 2019; Britannica, T. Editors, 2019).

This leads to the assumption that the supply and demand are directly impacting factors when analyzing the increased resale prices of sneakers. Taking this theory as a starting point, the following chapter is divided into research on the impact of the two variables supply and demand on price developments.

2.2.1 Supply

The initial suppliers in the sneaker market are the brands, such as adidas and Nike, which produce and release new sneakers. However, due to the resale market being of such major importance in the sneaker world, traditional consumer roles have shifted, enabling consumers to act not only as such, but additionally as suppliers in the resale market (Choi, 2017).

Releasing only small numbers of limited-edition sneakers is not uncommon to do for big brands. Adding to the scarce release, these brands oftentimes do not communicate to their consumers how many pairs a new release will comprise and choose exclusive sale locations

(Denny, 2020). Limiting the production and release of a certain product has been a common marketing strategy, increasing the value of both product and brand as consumers value those goods higher than easily attainable goods (Choi, 2017).

Lynn (1989) provides further reasoning to increased desirability resulting from scarcity of a good. In his research, he concludes that scarcity, may it be real or even just artificial, can be used to create a certain expectation of higher prices in customers. This is a result of people’s assumption of higher prices of scarce goods, hence a higher acceptance for those increased prices. Additionally, a higher price is often associated with higher quality, further enhancing the desirability of these items as status symbols. However, Lynn (1989) also mentions that the “assumed expensiveness of a scarce product should increase its status value only if the product category is used as a status symbol” (p. 272). Supported by research by Brock (1968) and Fromkin (1970), it can be assumed that scarce items are valued higher because the mere possession of such goods contributes to the owner’s feeling of personal uniqueness. Gierl & Huettl (2010) come to a similar conclusion, stating that for products used for conspicuous consumption, scarcity resulting from limited supply is advantageous in receiving improved product evaluations from consumers. Hwang et al. (2014), building on these findings, conclude that for some the price of a good is of little importance when given the chance to show wealth and social value by purchasing conspicuous and rare products. They further state that even manipulated, hence artificially made rarity, curbs the negative effect of price on desire to buy.

The fact that scarcity enhances the desirability of objects is best described on the basis of Commodity theory. This states that commodities that can be possessed, provide some form of usefulness to their possessors, but if they can be transferred from one possessor to another, they will be valued on the basis of their unavailability (Brock, 1968). Hence, the less available an object seems, the more a person values it.

Relating this back to the sneaker market, Choi (2017) suggests that knowing how scarce a new release will be, consumers will be attracted more to the attempt to get one of the limited pairs. Cassidy (2018) further supports this assumption, realizing a preference of consumers towards supply-scarce sneakers, which, as he states, is motivated by their consumption of sneakers as a conspicuous consumption product, meaning they can be used to display characteristics of the user.

Sneaker scarcity is not just fostered by brands only releasing a certain, small number of shoes, but also by selling these exclusively through certain stores and retailers. Additionally, in order to circumvent violence and cheating at releases, brands have opted to use online raffle systems, allowing users to try and "win" the chance to buy a pair online without having to queue in front of stores (Choi, 2017; Denny, 2020).

The scarcity of sneakers is however not solely related to limited supply by the brands. Sneakerheads participate no longer just as consumers in the market, but also as collectors and, most importantly for this thesis, as resellers. Noticing how valuable and profitable the resale of limited sneakers has become, professional resellers engage in creating so-called bots, hacking algorithms to enter the raffles and websites with the goal of buying as many pairs as possible (Choi, 2017; Denny, 2020). Researchers differentiate the behavior of sneakerheads in three types (Chu, 2013; Chu & Liao, 2007; Choi, 2017). First, sneakerheads who purchase sneakers with the intent of wearing, second, sneakerheads acquiring sneakers with the purpose of collecting, and lastly, sneakerheads buying sneakers with the intention to resell. However, the dividing lines between these three types of behaviors are not definite, as sneakerheads oftentimes show mixed behaviors (Choi, 2017). A sneakerhead might buy sneakers with the intent of wearing but might later decide to resell them.

Based on their motivation and behavior, sneaker resellers can be classified into three groups: consumer resellers, professional resellers, and mixed-role resellers. Consumer resellers sell sneakers that were acquired initially for their own use but resell for various reasons such as making a wrong purchase. Professional resellers on the other hand sell sneakers purchased with the motivation to resell and maximize profits and aim for a successful exchange in order to avoid high inventory levels. Finally, mixed-role resellers share behaviors and motivations of both consumer and professional resellers and acquire sneakers for both their own use and reselling (Chu & Liao, 2007; Choi, 2017).

The limited releases of sneakers together with more market participants who professionally resell lead to situations in which it is impossible for a sneaker fanatic to purchase the shoe on the primary market – website crashes, as well as bots hacking webpages to buy the sneakers, are not uncommon, forcing the unlucky sneakerhead to opt for resale markets and oftentimes pay a price way above the initial retail price. Additionally, some sneaker retailers engage in the practices of "back-dooring", a phrase referring to the possibility for resellers to buy large amount of stock at these stores before release date, hence before anyone else and especially individual sneaker fanatics have a chance to buy them authorized (Servantes, 2021).

After not being able to buy the desired object in the traditional way, the resale market allows the collectors and users a second chance of purchase. However, the question remains as to why the large price premiums come to place. The following paragraph hence highlights how demand of these sneakers ultimately contributes to an increased price in the resale market.

2.2.2 Demand

Another reason for increased prices in the resale market can be a demand too large to be satisfied by the already scarce supply.

Demand is driven by the customers. Neap & Celik (1999) state that the value the customer accounts to a product is made up of a cost paid together with a certain subjective valuation and an aspiration of receiving the product. Sweeney & Soutar (2001) describe value as a composition of both monetary but also personal, user specific attributes, which can change significantly when the owner changes. According to Sweeney & Soutar (2001), sneakers can hold the following different types of value: (1) functional value, quality and use, (2) emotional value, (3) social value, and (4) monetary value. Table 2.1 includes a description of the dimensions.

Table 2.1: Value dimensions of a sneaker

Emotional value	The utility derived from the feelings or affective states that a product generates
Social value (enhancement of social self-concept)	The utility derived from the product’s ability to enhance social self-concept
Functional value (price/value for money)	The utility derived from the product due to the reduction of its perceived short term and longer term costs
Functional value (performance/quality)	The utility derived from the perceived quality and expected performance of the product

Note. Adapted from “Consumer perceived value: The development of a multiple item scale” by Sweeney, J. C., & Soutar, G. N., 2001, *Journal of Retailing,* 77(2), p. 211.

The following presents research that explains how the at times seemingly immense demand of sneakers can be related to these types of value.

Functional Value Increase

The overall functionality of the sneakers does not increase during the period from retail to resale. However, seeing that sneakers are no longer solely desired for sports purposes but are also seen as a collectible (Semmelhack et al., 2015; Choi, 2017; Denny, 2020), the functional value of these is amended. Another theory that can lead to an increased functional value for consumers is the concept of availability heuristic. This describes the situation in which people will relate the easiest or most convenient explanation to a situation, e.g. relate Michael Jordan's exceptional achievements with the special shoes he was wearing while achieving this, rather than considering the years of hard work and training he had to go through for this (Tversky & Kahneman, 1973; Semmelhack et al., 2015; Choi, 2017).

Emotional Value increase

According to Choi (2017), "emotional value was derived when sneakers represented connections to others (friends, family, and endorsers) and connections to oneself (expression of one's identity)" (p. 96). An example for this can be celebrity endorsements, where "the image of a celebrity (e.g., talent, success, and excellence) is transferred to the particular brand or goods" (Choi, 2017, p. 38), which will in turn increase the value of the respective product or brand. This makes consumers feel connected with the respective celebrity, which can increase product value (Choi, 2017).

The success of this concept depends strongly on the endorser. The source credibility theory assumes that the effectiveness of a message is related to the expertness and trustworthiness of the source communicating the message. The source attractiveness model evaluates the effectiveness of the message based on the familiarity, likability and similarity of the source, which means that e.g. the criteria of how well a source is known determines how persuasive and attractive the source is (Hovland & Weiss, 1951; Seiler & Kucza, 2017).

These theories can be added to the Meaning Transfer theory by McCracken, which adds further explanation to the increase in value of goods when resold. According to McCracken (1986), items carry meaning which is transferred from the "culturally constituted world" (p. 71) to the consumer via a consumer good, facilitated through advertisements and fashion. Following his findings, goods offer a way to express culture and show belonging to a certain culture.

This research indicates that celebrity endorsements and collaborations increase the value of sneakers both at retail and at resale. The celebrity acts as a transmitter of cultural meaning on the product, which is then transferred to the customer who interprets that the cultural meanings from the respective celebrity are then contained also in the product.

Social Value Increase

Another dimension of value increase on the demand side is the increase due to social valuation. A lot of the sneaker cult revolves around the communities (see Chapter 1). The research shows that for members of certain groups and communities, it is of high importance to wear the "right" outfit, representing the group belonging and culture (Denny, 2020).

Sneakers can be described as luxury goods, meaning non-essential, rare and exclusive products (Tynan et al., 2010), that are used to display a certain social status (Hwang et al., 2014). The concept of conspicuous consumption, meaning the "spending on goods […] for the purpose of displaying wealth or social power or consumption deliberately intended to cause envy" (Veblen, 1889, as cited in Hwang et al., 2014, p. 1912), is seen as a way of showing wealth power, causing envy amongst peers and underpinning superiority of the owner (Veblen, 1889, as cited in Hwang et al., 2014). In the sneaker world, standing out and showing off the newest acquisition on social media or to the groups is part of the community life. Sneakerheads chase after rare releases to expand their private collection, but also to show their belonging to the groups (Denny, 2020).

The development of new customer groups, and the blurring borders between consumer and (re-)seller, have led to the increased significance of social connections amongst sneakerheads. Users even transform and sometimes create the new value of the product, seeing that oftentimes they are the ones offering the sneakers for resale.

Choi (2017) finds that sneakerheads overall assign social value to sneakers, as these give "a sense of social membership and social status to them" (p. 153). Denny (2020) supports this research, mentioning the relationship between wearing a certain style and membership of a certain subculture or tribe.

Another influencing factor on the social value of a sneaker is the hype around it generated through social media. Jacob A., a professional sneaker reseller, states that hype around a sneaker release is almost created by the sneaker culture itself (Maher, 2019). Choi et al. (2015) highlight the importance of social media, stating that sneakerheads oftentimes gather their purchasing information from unreliable sources, which are rumors or other leaked information. Increasing profit margins and communication thereof attract more market participants and shift the participant basis away from regular sneakerheads to those chasing profits (Cassidy, 2018).

Monetary Value Increase

Seeing that the typical consumer roles are shifting in the sneaker market, together with the development of artificial scarcity, exclusive locations of distribution and online raffle systems, the monetary value of sneakers is of increasing importance for sneakerheads (Choi, 2017; Denny, 2020).

The monetary value is influenced by the factors discussed above. Rarity increases the preference of consumers and hence their value perception, similarly to celebrity endorsement and collaborations.

Additionally, another reason customers are willing to pay such high prices is that according to Chu & Liao (2010), the awareness of an option to resell an item at a point later in time can enhance a customer's purchase intention. Hence, the preference of purchasing a certain product is impacted by the perceived chances of future resale. The better these perceived chances, hence, the larger the expected profit, the more willing a customer will be to purchase the item, even if it might not fit their personal taste (Chu & Liao, 2010).

The Hedonic Pricing Model

In order to analyze how the different features and value creators discussed above form a resale price for sneakers, the hedonic pricing model is a suitable measure (Ma & Treiber, 2020). The hedonic model defines the price of an object as the sum of the values of the characteristics of this object, which often finds application in the real estate market. Further research into this topic shows that e.g. the value of baseball cards is determined by factors such as popularity and success of the player, overall suggesting that cards showing famous players generally have high values (Mulligan & Grube, 2006).

Following the previous research, the resale value of sneakers is majorly impacted by its availability, collaborations and celebrities designing or wearing it, hype around the sneakers, its ability to increase the wearer's social status and other marketing features such as retro branding. One can therefore assume that additionally to the value that a piece of clothing holds for usage and technical features, sneakerheads may value their sneakers on the basis of emotions connected to memories related to these and the social values of being connected to the culture and community.

2.3 Machine Learning Techniques in the Context of Price Predictions

This chapter summarizes findings from the literature review on prior applications of machine learning techniques in the context of price predictions. First, findings on predicting prices in general are presented. This is followed by the applications in the context of predicting resale prices. Finally, findings on existing applications of machine learning techniques in predicting sneaker resale prices are presented.

2.3.1 Predicting Prices

The development and increased awareness of machine learning techniques allowed research scholars to enhance price prediction methods. An abundance of methods and models, each providing different advantages and disadvantages, has led to the proposal of many different approaches for price predictions.

Especially the usage of machine learning to predict commodity prices, e.g. of oil, gas, coal and agricultural products, has caught the attention of scholars. Additionally, the topic of forecasting energy and stock movements found resonance among analysts. Given the very limited pertinent research on algorithmic sneaker price prediction, the authors of this thesis extend the literature review to other industries and products, since machine learning for price prediction has been of wide interest for scholars across disciplines.

After comparing the predictive power of traditional forecast methods with machine learning methods such as *Neural Networks (NN)* and *Random Forest (RF)*, Herrera et al. (2019) come to the conclusion that the latter clearly outperform traditional econometric models and hence provide more accurate and stable forecasts for oil and gas prices.

When trying to predict and explain the Spanish day-ahead electricity prices, Díaz et al. (2019) chose a model using *Boosted Gradient Regressions Trees*, which not only provided them with stable predictions but also a way to analyze the impact the energy drivers have on the overall price. This allowed them to come to further conclusions about non-linear trends and relationships in their predictors.

Due to the importance of precise predictions and estimations of stock prices and other financial data, as well as the sometimes highly volatile price developments in financial markets, many researchers have focused on using machine learning methods to predict financial data. Chen et al. (2020) for example use statistical models such as *logistic regression* and *Linear Discriminant Analysis* as well as machine learning models like *RF, Extreme Gradient Boosting (XGBoost), Quadratic Discriminant Analysis* and *Support Vector Machines* to predict Bitcoin prices. They arrive at the conclusion that the statistical models are sufficient in complexity when predicting low-frequency data, while the more complex machine learning algorithms are better suited when analyzing high-frequency data. However, they point out the lack of sentiment in their analysis, aiming to provide future insights into the application of text mining as well as analyzing social network for better prediction results.

Another industry which has led researchers to incorporate machine learning techniques for reliable price predictions is the real estate and housing industry (Park & Bae, 2015; Truong et al., 2020; Rico-Juan & Taltavull de La Paz, 2021). The application of such methods leads to enhanced predictions of housing prices, going beyond the previously used hedonic-based models, and additionally contributes to cost savings when analyzing real estate (Park & Bae, 2015). Truong et al. (2020), after comparing various methods for house price prediction, come to the result that the different models each have different advantages and disadvantages. Evaluating their performance, they conclude that thanks to generalization, at times simpler regression methods perform better than more complex models such as *RF, XGBoost* and *Light Gradient Boosting Machine*. The main findings from this study are the necessity to make trade-offs between accuracy of models and computation time, and between simplicity and accuracy. Further, they highlight how often times, simplicity of models is best.

The application of *neural network algorithms* has sparked interest amongst researchers (Kohzadi et al., 1996; Adebiyi et al., 2014). However, many also mention the limitations these methods have, such as overfitting and parameter sensitivity (Pudaruth, 2014; Chai et al., 2019). After comparing a feed-forward back-propagation *Artificial Neural Network (ANN)* to a *RF* to predict energy consumption, Ahmad et al. (2017) conclude that the difference in performance is only marginal. Despite performing slightly better than *RF*, the authors highlight the advantages in tuning and handling the latter model and state the highly comparable predictive power of these two methods. This aspect is further taken up by Díaz et al. (2019) as well as Gaillard et al. (2016), who both point out the advantages the simpler regression trees have over *ANN,* these being simpler calibration and faster computation speed.

Moving away from the challenge of predicting a single mean value, many researchers have explored the predictions of intervals, using these results to generate scenarios (Nowotarski & Weron, 2014; Dvorkin et al., 2014).

2.3.2 Predicting Resale Prices

An area that has found many researchers applying machine learning methods for predicting resale prices is the car resale market (Lessmann & Voß, 2017; Pal et al., 2018; Pudaruth, 2014). Aiming to predict prices of used cars in Mauritius, Pudaruth (2014) uses four different machine learning methods *(Linear Regression, K-Nearest-Neighbors (KNN), Decision Trees* and *Naïve Bayes).* He bases the prediction of the values on several factors and attributes, such as age and make of the vehicle, technical details such as mileage and horsepower, but also includes information on interior style, paint color and additional features of the cars. He encountered problems with the predictive power of decision trees and naïve bayes methods, seeing that these are only able to handle output classes with numeric values, which forced him to classify the price attribute into ranges, leading to a loss of accuracy in his predictions (Pudaruth, 2014).

Pal et al. (2018), following a similar approach but using a *RF model*, also make use of car specific features such as age, make and origin, in order to predict the prices of used cars.

2.3.3 Predicting Sneaker Resale Prices

The currently available papers and research on predicting sneaker resale prices are of limited nature. The authors of this thesis would like to point out that there is a number of articles and reports released on websites such as *medium, github* and *Kaggle*, which give short insights into example usage of machine learning algorithms for resale price prediction (Zhang, 2020; Norman, 2020). Although the scope and insight these articles provide is of limited nature, they do highlight the increasing interest and importance of research of this topic.

Zhang (2020), using a neural network type, predicts prices on footwear with an error rate of ~30%. Norman (2020) compared the prediction results of a *RF* model, an *XGBoost* and a *Decision Tree Regressor* and concluded that the best performance was given by *RF*.

Shah (2019) in his thesis used an artificial neural network with data from StockX as well as Twitter data to predict whether a sneaker will resell for profit. Using data on the shoes, such as gender, brand and retail price from StockX, together with data from Twitter, which enabled him to quantify “hype”, his model had an accuracy range of 65-70%. This method combines the two areas hype and characteristics of a shoe to give a classification prediction of profit in relation to the retail price and average deadstock price (the price of new, unworn shoes on the platform).

2.4 Concluding the Literature and Hypotheses Formulation

Following the review of relevant literature, the authors identify several rooms for further research, which will be topic of this thesis. The authors realize that there is a lot of research available on price development as well as predicting prices, which shows that this is an interesting research topic overall. However, seeing the relatively new area of research, sneaker resale price prediction is not yet so advanced.

Based on the findings from demand and supply analysis together with the hedonic pricing model, the authors will include the importance of characteristics of sneakers in their price prediction and will hence attempt to predict the resale prices based on significant attributes such as color, silhouette and brand.

In order to facilitate the analysis of the research questions defined in Chapter 1.3, the authors formulated several hypotheses to evaluate in the course of this thesis. Following RQ1, the authors come to the following hypothesis in order to identify how the resale price is impacted by the characteristics of a sneaker:

H1: *The characteristics of a sneaker have different impact on the price development.*

This hypothesis is based on research by Ma & Treiber (2020), who find that the different characteristics such as brand and model differentiate sneakers. They apply the hedonic pricing model, which relates the value of an item to the sum of the values of the characteristics of this item. The authors of this thesis hypothesize that each variable, hence each feature of a sneaker, will have a different impact on the overall value.

H2: *The price increase on the resale market is related to the brand that released the sneaker in the first place.*

Brands and belonging are of major importance in the sneaker world (Semmelhack et al., 2015; Denny, 2020). The special relationship between brands and their consumers has been subject of many previous publications (Semmelhack et al., 2015; Denny, 2020; Choi, 2017). The high importance of belonging and wearing the “right” clothes (Choi, 2017) lead the authors to the hypothesis that the brand of a shoe is most of high importance when predicting the resale price of a sneaker.

In order to analyze in what way the resale value of a sneaker is impacted by the time of resell, the authors formulated the following hypothesis:

H3: *The resale value of a sneaker is at peak directly after release and decreases then.*

Ma & Treiber (2020) conclude that usually, the level of hype and communication about a certain sneaker is highest around its release date and will usually decline in the post release phase. Hence, resellers will often aim to resell the sneaker directly after the release date when hype and desire are still at peak (Ma & Treiber, 2020).

Following RQ2, and in order to address the different reseller needs in investment decisions, the authors come to the following hypothesis:

H4: *Complex models give significantly more accuracy in their predictions.*

This hypothesis comes from the author’s findings in Chapter 2.3, where more complex models, like RF or XGBoost, are widely applied in the context of price prediction and result in accurate predictions.

Following RQ3, the next hypothesis was formulated in order to thoroughly analyze the question of how collaborations influence the resale price.

H5: *Celebrity and other collaborations increase the resale price.*

Given the many different examples of sneaker brands making use of collaborations (e.g. adidas and Kanye West, Nike and Michael Jordan), which serve as a popular marketing resource and hence increase the desirability by inflicting more meaning to the item (McCracken, 1986), the authors assume higher resale value when a sneaker has been released in cooperation or collaboration with other brands or celebrities. This is assumed to be due to larger demand and hence even greater effect of artificial scarcity.

3. Methodology

The methodology section is divided in to two main parts. In the first part, the statistical learning models used in the study are presented. The second part focuses on the evaluation and validation methods.

3.1 Machine Learning Models

First, models for the regression setting are explained, including linear regression and tree-based regression models. This is followed by an overview of methods for the classification setting, consisting of logistic regression, KNN and classification trees.

3.1.1 Linear Regression

Linear regression is a relatively simple and widely applied approach to predict quantitative responses. Simple linear regression estimates a quantitative response Y built on a single predictor variable $X$ and presumes a linear relationship between $X$ and Y (James et al., 2013):

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

The error term ε represents a residual variable that accounts for the lack of fit with a model and is assumed to be independent of $X$ (James et al., 2013).

The simple linear regression model can be extended to fit multiple predictors by assigning each predictor a distinct slope coefficient in a single model. For p separate predictors, the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

The variable X_j represents the jth predictor whereas the regression coefficient β_j measures the relationship between that variable and the response and can be interpreted as the mean effect on Y of a single unit increase in X_j, while keeping all other predictors fixed. The least squares approach is a common approach to estimate the unknown regression coefficients, where $\beta_0, \beta_1, \ldots, \beta_p$ are selected by minimizing the residual sum of squares (RSS) (James et al., 2013):

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}\right)^2$$

With the obtained estimates for $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, predictions can be calculated with the formula (James et al., 2013):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + - \cdots + \hat{\beta}_p x_{ip}$$

3.1.2 Tree-Based Models

Regression Trees

In tree-based methods, the predictor space of the values for $X_1 . X_2, \dots, X_p$ is segmented into J unique and non-overlapping regions $R_1, R_2, \dots, R_J$. Observations falling into the region R_j are assumed to have the same prediction, which is the mean of the response values for the training set in R_j. The regions are divided so that the RSS is minimized, where $\hat{y}_{R_j}$ represents the mean of the training set within the jth region (James et al., 2013):

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2$$

Applying a top-down, greedy recursive binary splitting procedure can reduce complexity and increase computational efficiency. Starting at the top of the tree, the tree is split sequentially into two new branches at each node. The splits are selected in a greedy manner, meaning that the best split is chosen at that particular node without considering future splits down the tree. The predictor X_j and the cutoff point s are selected to split the predictor space into two regions $R_1(j, s) = \{X | X_j < s\}$ and $R_2(j, s) = \{X | X_j \geq s\}$ such that the RSS of the resulting tree is minimized (James et al., 2013):

$$\min_{j,s} \sum\nolimits_{i: x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum\nolimits_{i: x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2$$

Next, the two regions created are further split so that the RSS is minimized. This process is repeated until a previously determined stopping criterion is reached. After the regions $R_1, \dots, R_j$ have been identified, the mean of the training set of a region to which a test observation belongs to represents the response prediction (James et al., 2013).

Bagging, Random Forest and Boosting

Tree-based methods are easy to interpret, but compared to other supervised learning methods, they tend to deliver poorer results in terms of prediction accuracy. Moreover, trees tend to be non-robust and suffer from high variance. A small change in the training data used to fit a tree can lead to significantly different results. However, the variance can be reduced by using methods like bagging, RF or boosting, where multiple trees are fitted and then combined to yield a single prediction. These methods can also improve the prediction accuracy significantly (James et al., 2013), and will be introduced in this section.

Bagging

In the bagging procedure, samples from the training set are repeatably taken to generate B different bootstrapped training sets. A tree is fit for each bth training set to generate a prediction. Finally, all resulting predictions $\hat{f}^{*b}(x)$ are averaged to obtain a final prediction (James et al., 2013):

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$$

Although the individual trees have high variance, they result in low bias. So, by taking the average of the trees the variance can be decreased. However, in case of a dominant predictor in the data set, most of the fitted trees might choose the same predictor for the initial split, leading to highly correlated predictions (James et al., 2013).

Random Forests

Similar to bagging, the RF approach constructs multiple decision trees on bootstrapped training samples. However, at each split in the tree, only a random sample of m predictors are considered. By forcing each split to consider a subset of the predictors, other predictors apart from the dominant predictors are given a chance to be chosen for the initial split. This approach can increase the reliability of the resulting trees by decorrelating the trees and making them less variable (James et al., 2013).

Boosting

Boosting is another method to increase the prediction accuracy of decision trees. Like in bagging, multiple trees are constructed, however, the trees are grown sequentially meaning that each tree considers information from already fitted trees. In contrast to bagging, boosting does not take bootstrapped samples of the training set, the trees are instead fit on a modified version of the original data set. The procedure includes fitting and combining many trees $\hat{f}^1, \ldots, \hat{f}^B$. Instead of fitting the model using the outcome Y, the residuals of the current model are used as the response to construct a decision tree. This model is then added to the fitted function so that the residuals are updated. The algorithm is described in Table 3.1 (James et al., 2013).

Table 3.1: Boosting algorithm for regression trees

```
1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all $i$ in the training set.
2. For $b = 1,2, \ldots, B$, repeat:
    a. Fit a tree $\hat{f}^b$ with $d$ splits ($d + 1$ terminal nodes) to the training data $(X, r)$.
    b. Update $\hat{f}$ by adding in a shrunken version of the new tree:
           $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$
    c. Update the residuals,
           $r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$
3. Output the boosted model,
           $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$
```

Note. Adapted from *An introduction to statistical learning: with applications in R* (8th ed., p.323) by James, G., Witten, D., Hastie, T., & Tibshirani, R., 2013, Springer.

Generally, statistical learning models that are trained slowly tend to perform well. In boosting, selecting a small number of splits results in smaller trees that learn slowly. This can be adjusted with the parameter d. The training process can further be slowed down by adjusting the shrinkage parameter λ, which will allow the construction of more and different shaped tress to handle the residuals (James et al., 2013).

3.1.3 Classification Models

The regression models discussed so far assume the response variable to be quantitative. However, in many scenarios in this study, the response variable is qualitative. Thus, this section introduces models for the classification setting.

Logistic Regression

Logistic regression is based on linear models and is a simple model for classification problems with two outcomes. It models the probability that a qualitative response Y belongs to a specific category, that is $p(X) = \Pr(Y = 1| X)$. In order to obtain outputs between 0 and 1 for all values of X, the logistic function is used to model $p(X)$ (James et al., 2013):

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

The maximum likelihood method can be used to estimate the regression coefficients β_0 and β_1. The estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen such that the likelihood function is maximized (James et al., 2013).:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

Finally, predictions can be made by plugging the coefficient estimates into the model function $p(X)$. The obtained estimated probability is then classified into categories based on a set threshold (e.g. 0.5) (James et al., 2013).

The simple logistic regression model can be extended to fit multiple predictors. For p separate predictors, the multiple logistic regression model takes the form as follows (James et al., 2013):

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Similarly to simple logistic regression, the maximum likelihood method can be used to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_p$.

K-Nearest Neighbors Classification

KNN is a simple model that classifies observations based on a similarity measure, e.g. the Euclidean distance. More specifically, the KNN classifier finds K observations in the training set that are the closest to the test observation x_0. The identified closest observations are represented by $\mathcal{N}_0$. Next, the probability for class j is estimated as the fraction of observations in $\mathcal{N}_0$ whose response observations equal j. The test observation x_0 is then predicted to belong to the class with the highest probability (James et al., 2013):

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

The choice of K effects the KNN classifier significantly, and as K varies, there is a bias-variance tradeoff. A flexible classifier with a small value for K can result in low bias but very high variance. On the other hand, a less flexible classifier with a higher value for K can lead to low variance but high bias. Hence, it is critical to select the correct level of flexibility in order the obtain a good classifier (James et al., 2013).

Classification Trees

Classification trees predict that an observation belongs to the most frequently occurring class of training data in the observation's region. Similarly to regression tress, classification trees are built using recursive binary splitting. As the RSS is not suitable for the classification setting, the classification error rate can be used as an alternative criterion for splitting. The classification error rate is the proportion of training observations in a region that do not belong to the most frequently occurring class. With $\hat{p}_{mk}$ representing the fraction of training observations in the mth region that are from the kth class, the classification error is defined by (James et al., 2013):

$$E = 1 - \max_k(\hat{p}_{mk})$$

In practice, the Gini index and the cross-entropy are preferred over the classification error to evaluate the quality of a particular split. These measures are differentiable, and hence more suitable for numerical optimization (Hastie et al., 2009; James et al., 2013). The Gini index measures the total variance across K classes, and is given by (James et al., 2013):

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \left(1 - \hat{p}_{mk}\right)$$

The Gini index takes on a small value for nodes containing mostly observations belonging to a single class (node purity). The alternative measure, cross-entropy is defined by (James et al., 2013):

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

Similarly, the entropy takes on a small value for pure nodes. As the Gini index and entropy are more sensitive to node purity, they are preferred for tree pruning. However, if the focus is on prediction accuracy of a pruned tree, the classification error is favored (James et al., 2013).

3.2 Assessing Model Accuracy and Validation

This section explains the various performance measures used to evaluate the accuracy of both regression and classification models. Finally, the method of validation is presented.

3.2.1 Assessing Model Accuracy

Mean Squared Error

Measuring the match between predictions and the observed data allows to evaluate the performance of a statistical model. In detail, the closeness of the predicted value $\hat{f}(x_i)$ to the true value y_i for the ith observation must be quantified. A common measure for regression models is the *root mean squared error (RMSE).* This measure is a function of a model's residuals and is calculated by taking the square root of a model's *mean squared error (MSE),* which is another common measure for regression models (Kuhn et al., 2016). The MSE is calculated as follows (James et al., 2013):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

A small MSE indicates that the predictions are close to the true values whereas a large MSE indicates that the predictions diverge significantly from the true values. More accurately, this MSE can be referred to as the training MSE, as it is calculated based on a model fitted on the training data. However, the test MSE, that is the accuracy of the estimates from a model on previously unseen test data, is more interesting. A low training MSE does not guarantee a low test MSE, especially in the case of more flexible models. Increasing the flexibility of a model leads to curves that follow the training data more closely, and hence, the training MSE declines. This may lead to overfitting of the data, where a model results in a small training MSE, but performs poorly on the test set and thus results in a large test MSE. Overfitting is caused by the model following the training set too closely and picking up random patterns which are not present in the unknown, true function f. So, rather than a model that yields to the lowest training MSE, a model that results in the lowest test MSE should be selected. The test MSE is computed by taking the average squared prediction error for the previously unseen test observations (x_0, y_0) (James et al., 2013):

$$Ave\left(y_0 - \hat{f}(x_0)\right)^2$$

These insights apply accordingly to the measure RMSE: the smaller the RMSE on the test set, the better the model.

The Bias-Variance Trade-Off

The expected test MSE for a value x_0 can be decomposed in the variance of $\hat{f}(x_0)$, the squared bias of $\hat{f}(x_0)$ and the variance of the error terms ϵ (James et al., 2013):

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 - Var(\epsilon)$$

The expected test MSE $E\left(y_0 - \hat{f}(x_0)\right)^2$ is obtained by estimating f repeatedly on multiple training sets, testing each at x_0, and calculating the average over all potential values of x_0 in the test set. A model that results in both low variance and low bias will minimize the expected test error. The variance of a model describes how much $\hat{f}$ changes if the model is trained on different portions of the training data. In a more flexible model with high variance, little changes in the training set can lead to major changes in $\hat{f}$. The bias on the other hand describes the error caused by approximating a complex real-life scenario by a simple or restrictive model.

Typically, as the flexibility of a model is increased, the variance grows and the bias declines. Generally, the bias decreases at a relative higher rate initially than the variance increases, and hence, the expected test MSE decreases. But at a certain point, increasing the flexibility of a model has only a small effect on the bias, and begins to increase the variance considerably. In this case, the test MSE increases. This relationship is referred to as the *bias-variance trade-off*, and a model with a good test performance is one that results in a low variance as well as a low squared bias (James et al., 2013).

Residual Standard Error and R^2 statistic

The *residual standard error (RSE)* can be computed in order to quantify to which extent a model fits the data (James et al., 2013):

$$RSE = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The RSE estimates the standard deviation of the residuals of a model, hence it can be regarded as a measure of lack of fit. A small value for the RSE indicates that the model results in predictions that are close to the true values, and the model fits the data well. On the contrary, a large RSE indicates that the predictions are far from the true observations, and the model fits the data poorly. However, as the RSE is measured in the units of Y, it might be difficult to assess what a good value for the RSE is (James et al., 2013).

The R^2 *statistic* is an alternative measure to quantify the model fit. Unlike the RSE, the R^2 statistic is scale-invariant and takes on a value between 0 and 1, making the interpretation of the measure clearer. The R^2 statistic is computed as followed (James et al., 2013):

$$R^2 = 1 - \frac{RSS}{TSS}$$

RSS is defined in Chapter 3.1.1 and quantifies the total unexplained variability in Y after a model is run. $TSS = \sum(y_i - \bar{y})^2$ is the total sum of squares (TSS) and can be interpreted as the total variability built in Y before a model is run. So, the R^2 statistic represents the fraction of variability in the dependent variable Y that is explained by X. A value close to 1 suggests that the model explains a large proportion of the variability, and hence, that the model fits the data well (James et al., 2013).

Including more independent variables in a model will always increase the R^2 statistic, as it allows the model to fit the training data more accurately (James et al., 2013). The *adjusted* R^2 *statistic* is a modified version of the R^2 statistic where n represents the number of data points and p represents the number of independent variables:

$$adjusted\ R^2 = 1 - \frac{(1-R^2)(n-1)}{(n-p-1)}$$

It adjusts for the number of independent variables in a regression model and shows whether the additional variable improves model performance (James et al., 2013).

Assessing Classification Models

Many models for regression can also be applied in a classification setting, however, the metrics for performance measurement of regression models are not applicable in the classification setting. Thus, the evaluation method differs for classification.

Confusion Matrix

The *confusion matrix* is commonly used to evaluate the performance of a classification model. The observed and predicted classes are displayed in a matrix table. For a two-class problem, the confusion matrix results in four combinations of observed and predicted values. Table 3.2 shows an example. The two diagonal cells of the matrix, labeled as *true negatives (TN)* and *true positives (TP),* represent cases where the class predictions were correct. The two off-diagonal cells, labeled as *false positives (FP)* and *false negatives (FN)*, refer to cases where the class predictions were false (Kuhn et al., 2016).

Table 3.2: Confusion matrix

		Observed Values	
		Negative (0)	Positive (1)
Predicted Values	Negative (0)	TN	FN
	Positive (1)	FP	TP

Metrics

From the confusion matrix, the overall *accuracy rate (ACC)* and the *error rate (ERR)* can be derived to assess the model's performance:

$$ACC = \frac{TN + TP}{TN + FN + FP + TP}$$

$$ERR = \frac{FP + FN}{TN + FN + FP + TP} = ACC - 1$$

The error cost of positives and negatives can differ based on the purpose of the model. Hence, more informative metrics and can also be derived from the confusion matrix. By dividing the predictions by the observed classes, the corresponding true and false rates can be calculated. The two true rates, *true positive rate (TPR)* and *true negative rate (TNR),* are calculated as follows:

$$TPR = \frac{TP}{TP + FN} \quad , \quad TNR = \frac{TN}{TN + FP}$$

The two false rates*, false positive rate (FPR)* and *false negative rate (FNR),* can be calculated similarly:

$$FPR = \frac{FP}{FP + TN} \quad , \quad FNR = \frac{FN}{FN + TP}$$

If the level of accuracy for a model is assumed to be fixed, a trade-off between TPR and TNR, also called sensitivity and specificity, occurs. To evaluate this trade-off, the *Receiver Operating Characteristic (ROC)* curve can be used which will be presented next.

ROC curve and AUC

The ROC curve plots the TPR against the FPR (1-TNR) over all possible thresholds (see Figure 3.1). An optimal model, where the TPR equals 1 and the FPR equals 0, would result in a ROC curve that hugs the top left corner (Kuhn et al., 2016). Since ROC curves consider all possible thresholds, they are suitable for comparison of various classifiers. *The area under the ROC curve (AUC)* is a metric for measuring a classifier's performance over all possible thresholds. The higher the value of AUC, the better the classifier performs, resulting in a ROC curve that hugs the top left corner (James et al., 2013).

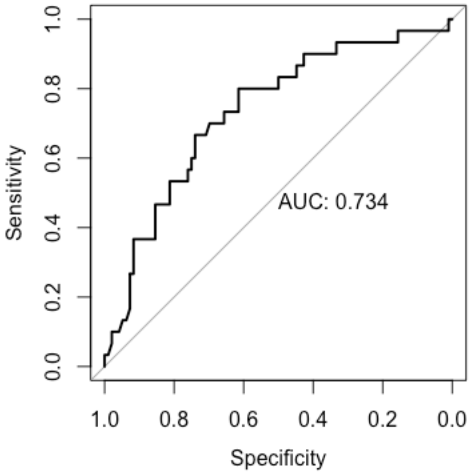


Figure 3.1: ROC curve

3.2.2 Validation

Validation can be used to assess a model's performance. As described in Chapter 3.2.1, training and validating a model with the same data sample may lead to overfitting. Hence, a fitted model should be assessed based on how well it performs when predicting previously unseen data. The *Validation Set Approach* is one method to achieve this, where the data is split randomly into a training set and a testing set. The training set is used to fit the model, and the response observations in the testing set are predicted using the fitted model in order to estimate its performance on previously unseen observations (James et al., 2013).

The test error estimated with the validation set approach can be very variable, depending on the random train/test split. *Cross-validation (CV)* methods where the validation process is repeated multiple times and the mean of all test errors is calculated, deal with this problem. One method for CV is *k-Fold Cross-Validation*. Here, the data set is divided into k groups, where the first group acts as a validation set and the remaining $k - 1$ groups are used to fit the model. This process is repeated k times, where a different group acts as a validation set each time, and the test MSE_i is calculated. The resulting k estimations of the test error rate, $MSE_1, \dots, MSE_k$, are then averaged to calculate the k-fold CV estimation (James et al., 2013):

$$CV_k = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

The choice of k influences the bias-variance trade-off. In practice, usually a 5-fold CV or 10-fold CV is chosen, as these empirically result in test error rates with a good balance between bias and variance (James et al., 2013). Due to the larger dataset and the high number of models fit in this study, the authors decided to perform a 5-fold CV on all models to save processing time.

4. Data and Preprocessing

This chapter consists of four sections. First, the dataset used in this thesis is introduced. This is followed by a description of the data preprocessing steps, further followed by a walk-through of the feature engineering steps. Lastly, an overview of the final dataset is given.

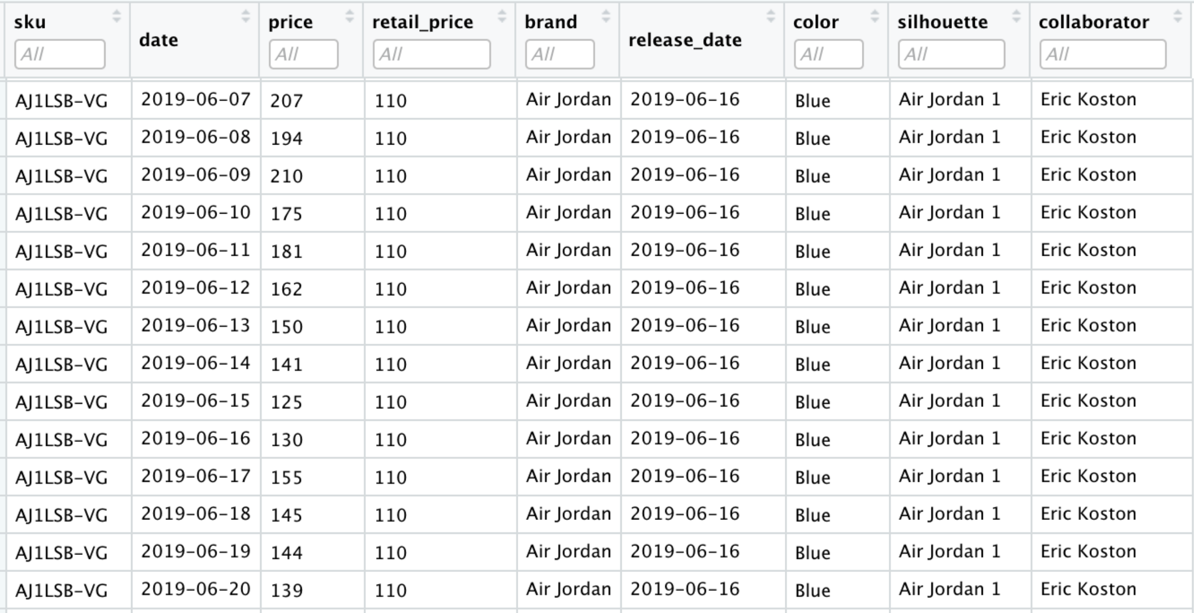
4.1 Introduction to the Dataset

The dataset was retrieved from StockX.com. StockX is the leading marketplace for sneaker resale and provides detailed historical resale price trends on sneakers sold over the platform. This includes only data on deadstock sneakers. The retrieved dataset from StockX consists of daily average resale ***price*** in USD of 728 sneakers. For the following analysis, the data used was limited to the top 3 reselling brands on StockX – adidas, Nike and Air Jordan – and included only the most popular sneakers which were released in 2019. "Most popular sneakers" are those that have been resold most frequently on StockX.com. This ensured that the daily average resale price of a sneaker was calculated on a large enough sample size (>= 200).

Additionally to the daily average resale prices per sneaker and ***date*** on which this resale price was paid, the original ***retail price*** as well as the ***stock-keeping-unit (sku)*** code of each sneaker were obtained from StockX.com.

The dataset was further enriched by including details obtained from GOAT.com, another popular marketplace for sneakers and sneakerheads. Additional details of the sneakers - ***release date, brand, color, silhouette*** and ***collaborator*** information - were gained from GOAT, where this data is provided in a very structured layout.

The initial dataset consisted of 276,932 rows and 9 columns. Figure 4.1 shows a snapshot of the retrieved dataset before any preprocessing and feature engineering steps.



sku	date	price	retail_price	brand	release_date	color	silhouette	collaborator
AJ1LSB-VG	2019-06-07	207	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-08	194	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-09	210	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-10	175	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-11	181	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-12	162	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-13	150	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-14	141	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-15	125	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-16	130	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-17	155	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-18	145	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-19	144	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston
AJ1LSB-VG	2019-06-20	139	110	Air Jordan	2019-06-16	Blue	Air Jordan 1	Eric Koston

Figure 4.1: Snapshot of the dataset retrieved from StockX and GOAT

4.2 Dataset Preprocessing

In a first step, some observations in the dataset were discarded. StockX provides information on resale prices even before the release date of a sneaker. Because these usually refer to even more limited sneakers, and the sneakers are usually obtained via unofficial trade routes and personal relations (e.g. "backdooring"), these were not considered in this analysis, and hence, the dataset was filtered to only include price data for days on and after the release date. Further, in some cases, the same sneaker had more than one price allocated on one date, which is due to the registering of different times in a day on StockX. In these cases, the earlier price entry was kept.

Furthermore, the data extracted was limited to a time period of 180 days, roughly 6 months, after a shoe's release. This allowed for a sufficient amount of data for the modelling, while keeping the processing time manageable.

Next, all columns containing date and time data were set to be seen as dates. In order to gain an understanding of price developments after release, a variable ***days_after_release*** was added which indicates the number of days after release for each row, hence giving an information of price development after release per sneaker.

Finally, for better analysis of price development, and to give more options for modelling, a column, ***p_premium***, containing the price premium per entry was added. The price premium was calculated as follows:

$$price\ premium = \frac{price - retail\ price}{retail\ price} * 100$$

4.3 Feature Engineering

The original dataset included a total of four categorical variables. These were *brand, collaborator, silhouette,* and *color.* The one-hot-encoding method is an in practice commonly used approach to deal with categorical variables in datasets. This is a method in which new dummy variables are introduced that replace the categorical variable and take the value 0 or 1. A new column is introduced for each category. For instance, the ***brand*** feature has three possible categories: "Nike", "adidas" and "Air Jordan". To encode these, three new features are created, and a feature is set to equal 1, if the brand for this data point has the corresponding value, and 0 otherwise. So, for each data point, only one of the three features will equal 1. The original brand column is dropped and only the three new binary features are kept. The result were three new columns named *brandNike, brandadidas* and *brandAirJordan*.

This works well for those categorical variables, which do not have too many levels, as a large number of levels would add a large number of columns to the data set. Hence, for the variable *brand*, which only had three levels, the method could be applied very well. However, for those variables which have a large number of levels, it is difficult to apply this method. Hence, for example for the categorical variable ***collaborator***, the authors decided to only use binary encoding on the question whether or not the sneaker was released in collaboration, but not identify the individual collaborators. Seeing that there were 76 levels in this variable, this would have significantly increased the size of the dataset, which would have resulted in losses in efficiency. So, a binary variable, ***collab_b***, was added to the dataset, indicating whether a certain sneaker was released in collaboration. This variable is set "0" if without, and "1" if released in collaboration with another designer, celebrity or brand.

The next categorical variable, ***silhouette,*** describes the form or model of a sneaker. A first overview of the data showed that this variable had over 120 different levels. Instead of keeping all of these, and hence substantially increasing the size of the data set when one-hot encoding, the authors decided to keep the most important ones, such as Air Jordan 1, and combining those silhouettes that are very similar (e.g. Air Max 1 and Air Max 90). Finally, those silhouettes, that did not fit any grouping or that did not appear frequently, were added to the "other" category, respectively of their brand (e.g. *otheradidas*, *otherNike* or *otherAirJordan*). Resulting from this, *silhouette* had only 10 levels left, which could then easily be encoded using the one-hot encoding described above.

The last categorical variable, ***color,*** consisted of 18 levels, where many of these did not appear very frequently in the data. Again, to make the data more manageable, the authors decided to group colors from the same type (e.g. add "teal" to "blue"), which resulted in *color* only having 6 levels. This also allowed for easy encoding following the one-hot encoding mechanism and resulted in 6 new columns.

4.4 Final Dataset

After all pre-processing activities as described above, the final dataset consisted of 123,405 observations of 31 columns. Table 4.1 summarizes the preprocessing steps and all columns in the final dataset.

Table 4.1: Overview of columns in the final dataset

Name of variable	**Description/definition**	**Preprocessing steps**	**Data type**
price	Resale value of the sneaker on one exact day	None	Integer
retail_price	Initial retail price of the sneaker, as set by the brand	None	Integer
brand	Name of the brand releasing the sneaker (Adidas, Nike and Air Jordan)	One-hot encoding, hence three dummy columns	Factor, then binary after encoding
color	Color of the sneaker, originally in 18 levels	Grouping into 6 colors, which were then encoded using one-hot encoding	Factor, then binary after encoding

silhouette	Silhouette, hence, shape/model of the sneaker, originally in 129 levels	Grouping into 10 silhouettes, which were then encoded using one-hot encoding	Factor, then binary after encoding
collab_b	Dummy variable indicating whether or not a shoe was released in collaboration	Dummy encoding, 0 = no collaboration, 1 = collaboration	Binary
days_after_release	Difference between release date and date of the resale price (sounds weird)	Made as numeric	“Difftime”, numeric
p_premium	Premium at which the resale price lies in comparison to the original retail price	Percentage	Numeric

5. Approach, Modelling and Results

This chapter includes three main parts. The first section describes the authors' approach in exploring RQ2. The second section shortly summarizes the authors' decision on the machine learning methods selected for modelling. Finally, the modelling process and results are presented in detail in section three.

5.1 Approach

The following sections describe the authors' approach of creating two reseller personas in order to explore RQ2 in more depth. Based on the author's assumptions made on the two personas, exemplary decisions that may arise during the two personas' reselling activities are identified.

5.1.1 Creating Two Reseller Personas

Based on the reseller classification in Chapter 2.1.1, two reseller personas will be created in order to explore research questions RQ2a and RQ2b. The focus will be on professional resellers and mixed-role resellers, as these groups tend to engage more in resale activities than consumer resellers. Table 5.1 summarizes the assumptions made by the authors in order to create the two personas. A professional reseller's primary motivation in purchasing a sneaker is to maximize profits by reselling, that is, reselling a sneaker at a higher price than initially paid. As professional resellers want to avoid high inventory levels, the authors assume the secondary motivation to be making a successful transaction. On the other hand, mixed-role resellers are presumed to purchase sneakers with the primary motivation of collecting, and with the secondary motivation of selling at a profit or minimizing loss. For the purposes of simplicity and comparability, the authors assume that both personas acquire and resell new, unworn pairs of sneakers (referred to as "deadstock" in sneaker terminology). This assumption also arises from the fact that the site Stockx.com, where the data was retrieved from, only lists deadstock sneakers.

Based on the previous literature review, professional resellers tend to utilize bots or engage in "backdoor" sales to secure limited sneakers. Hence, the authors assume that the professional reseller acquires a sneaker on the release date at retail price. Unable to compete against bots, mixed-role resellers tend to turn to the secondary market. Therefore, the authors assume that the mixed-role reseller purchases a sneaker on the secondary market at the current resale price. Further, the authors assume that the mixed-role reseller aims to purchase a sneaker within 30 days after its release for their collection. Regarding the time of resell, the authors assume that professional resellers aim to resell within 14 days after a sneaker's release due to limited storage. On the other hand, mixed-role resellers are presumed to keep a sneaker in their collection for a longer period of time and resell 180 days after a sneaker's release.

Table 5.1: Assumptions made to create the reseller personas

	Professional Reseller	**Mixed-Role Reseller**
Primary motivation	Maximizing profit	Collecting sneakers
Secondary motivation	Making a successful transaction to keep inventory time and cost down	Selling at a profit or minimizing loss
Sneaker condition (at the time of purchase and time of resell)	New, unworn pair ("deadstock")	New, unworn pair ("deadstock")
Time of purchase	On the release date	Within 30 days after release
Purchase price	Pays retail price	Pays current resale price
Time of resell	Within 14 days after release due to limited storage space and costs of holding inventory	On day 180 after release

5.1.2 Persona Decisions and Scenarios

In order to analyze and explore the RQs in more depth, to show the applicability of machine learning methods to the overall aim of predicting sneaker resale prices, and to show how these can support investment decisions, the following analysis was conducted on the basis of exemplary questions that may arise for the two different personas.

The two different personas (see Table 5.1) display different motivations to engage in sneaker (re-)sale, and out of these different motivations, different questions arise regarding their personal investment activities and decisions.

Professional resellers e.g. need to make buying decisions based on the possibility of quickly reselling the shoe at a profit, as their overall aim is profit maximization while keeping inventory costs at a minimum. Questions to consider are hence how the price premium of a sneaker will develop in the time shortly after release, what price can be reached after a few weeks following release, and if a certain sneaker will reach a certain (positive) price premium. Additionally, seeing that reselling quickly is of key importance for the professional reseller, information on when to best resale a certain shoe is also of relevance.

Mixed-role resellers have collecting sneakers as their main priority, hence the maximization of price or price premium is of less importance. Questions that can arise for these types of market participants would be when to buy a sneaker, if for example the aim would be to resell it at +-0 after a longer holding period of 6 months. Also, the development of the resale price over a longer time frame is important, since the mixed-role reseller might be interested in an estimate of long-term price development to resale after having had a shoe in their collection. Additionally, a mixed-role reseller, who is interested in purchasing a sneaker for their collection, might be interested in gaining information on when a sneakers' resale price is closest to its retail price, hence is resold at a low price premium.

These decisions are summarized in Table 5.2 for the professional reseller, and in Table 5.3 for the mixed-role reseller. The tables also include the researcher's approach in analyzing the two personas' questions. In the following, these questions and respective decisions to be made will be investigated by applying different machine learning methods to the data.

Table 5.2: Questions and scenarios of the professional reseller

Questions of the Professional Reseller	**Approach and Scenarios**
Q1: Should the reseller purchase a sneaker at retail price on the release date if they want to achieve a price premium of at least X% within 14 days?	The aim was to predict the binary variable *buy* which is defined as 1 if the achievable price premium for a sneaker within 14 days after its release is greater than 0%/15%/25%/50%, and 0 otherwise. ● **Scenario 1.1:** The reseller wants to achieve a price premium greater than 0% within 14 days. ● **Scenario 1.2:** The reseller wants to achieve a price premium greater than 15% within 14 days

	● **Scenario 1.3:** The reseller wants to achieve a price premium greater than 25% within 14 days. ● **Scenario 1.4:** The reseller wants to achieve a price premium greater than 50% within 14 days.
Q2: How does the resale price of a sneaker develop in the first 14 days after its release?	In order to analyze this question, the variable *price*, that is the current resale price of a sneaker, for the first 14 days after a sneaker's release was predicted.
Q3: What is the maximum price premium that the reseller can achieve within 14 days after a sneaker's release?	In order to analyze this question, the variable *max_p_premium*, that is the maximum value for the price premium of a sneaker within 14 days after its release, was predicted.
Q4: When should the reseller sell a sneaker if they aim to achieve a high price premium?	The aim was to predict the binary variable *sell* which is defined as 1 for days where the price premium of a sneaker is highest/within the 3 highest values/within the 5 highest values, and 0 otherwise. ● **Scenario 4.1:** The reseller is willing to sell on the day where the price premium is the highest. ● **Scenario 4.2:** The reseller is willing to sell on days where the price premium is within the 3 highest values. ● **Scenario 4.3:** The reseller is willing to sell on days where the price premium is within the 5 highest values.

Table 5.3: Questions and scenarios of the mixed-role reseller

Questions of the Mixed-Role Reseller	**Approach and Scenarios**
Q5: Should the reseller purchase a sneaker at the current resale price if they want to achieve a price premium of at least X% on day 180 after release?	The aim was to predict the binary variable *buy* which is defined as 1 if the achievable price premium for a sneaker on day 180 after its release is greater than 5%/0%/-5%, and 0 otherwise. ● **Scenario 5.1:** The reseller is willing to buy if they can expect to achieve a price premium of at least 5% on day 180 after release. ● **Scenario 5.2:** The reseller is willing to buy if they can expect to achieve a price premium

	of at least 0% on day 180 after release. ● **Scenario 5.3:** The reseller is willing to buy if they can expect to achieve a price premium of at least -5% on day 180 after release.
Q6: How does the average monthly resale price of a sneaker develop in the first 6 months after its release?	In order to analyze this question, the monthly average of the variable *price*, that is the current resale price of a sneaker, for the first 6 months after a sneaker's release was predicted.
Q7: When should the reseller buy a sneaker if they aim to pay a low price premium for it?	The aim was to predict the binary variable *buy* which is defined as 1 for days where the price premium of a sneaker is lowest/within the 3 lowest values/within the 5 lowest values, and 0 otherwise ● **Scenario 7.1:** The reseller is willing to buy on the day where the price premium is the lowest. ● **Scenario 7.2:** The reseller is willing to buy on days where the price premium is within the 3 lowest values. ● **Scenario 7.3:** The reseller is willing to buy on days where the price premium is within the 5 lowest values.

5.2 Model Selection

The questions and respective scenarios of the two personas were investigated by applying different machine learning methods to the data. Considering the various assumptions made regarding the two personas, their expectations towards a predictive model may also differ. As the professional reseller is presumed to be living off the sneaker business, the authors make the assumption that this persona is primarily interested in a model that delivers accurate predictions. Hence, the professional persona is assumed to accept more complex models that may be more difficult to interpret but lead to highly accurate predictions. As opposed to the professional reseller, the mixed-role reseller is assumed to be primarily interested in simpler predictive models that are easy to apply and interpret, while also giving fairly good predictions. However, as opposed to the professional reseller, the very best and most precise results are not required as the mixed-role persona is not primarily living off the sneaker business.

Based on these assumptions, the authors decided to include both simpler models that are easy to interpret, and more complex models that may lead to more accurate predictions. Below is an overview of the selected methods.

Linear Regression and KNN

As described in Chapter 3.1.1, *linear regression* is a relatively simple and widely applied method to predict quantitative responses. Linear effects on the response variables are easy to quantify and describe, hence in general, linear regression yields interpretable models. Similarly, *KNN* provides another simple method to fit. Hence, KNN and linear regression models were fitted in order to investigate the questions with quantitative responses, that is, Question 2 and Question 3 of the professional reseller, and Question 6 of the mixed-role reseller.

Simple Classification Models

To investigate the questions that require predicting qualitative responses, the authors decided to include two simple classification models. Both *logistic regression* and *KNN classification* methods are relatively easy to interpret and are described in detail in Chapter 3.1.3. Especially for the mixed-role reseller, who is primarily interested in the interpretability of a model, these two methods have the potential to offer simple and practical approaches for evaluating buy/sell decisions.

Tree-Based Methods

The authors decided to include various tree-based methods for both regression and classification settings. As described in Chapter 3.1.3, *simple decision trees* are relatively easy to interpret, and hence, they are potentially of interest for the mixed-role reseller. As simple decision trees tend to result in poorer prediction accuracy, the authors decided to also include two methods that fit multiple trees, namely *RF* and *XGBoost*. Compared to simple tree models, these models are more complex, but can reduce the variance and thus, increase the accuracy of predictions. Further, the literature review (see Chapter 2.3) showed that both RF and XGBoost are widely applied in the context of price predictions and result in accurate predictions. Hence, they are especially interesting for the professional reseller who primarily wants accurate predictions.

5.3 Modelling and Results

This section is divided into three main parts. In the first part, the previously mentioned machine learning methods are applied, and their performance assessed for each scenario of the professional reseller. The same is done for the scenarios of the mixed-role reseller in the second part. Finally, the main findings from the analysis are summarized in part three.

5.3.1 Professional Reseller

As described in Chapter 5.2, the professional reseller is assumed to be interested most in high accuracy. In order to provide this, the authors used linear regression, logistic regression, KNN, RF and XGBoost to provide models to fit Questions 1-4. The modelling process is described in the next section. This is followed by the analysis of the results.

Modelling

Fitting Linear Regression, Logistic Regression and KNN

The authors fit the models using the package *caret*. For the regression scenarios, linear models serve as a simple starting point, as no tuning is needed since the linear regression method has no hyperparameters. The authors decided against transforming the predictors to address a potential non-linearity in the data. The reasoning behind this decision was to keep the simplicity and easy interpretability of the models, as the linear model was primarily chosen by the authors to accommodate the needs of the mixed-role-reseller.

For the classification scenarios, the authors chose logistic regression as a simple starting model. There was a class-imbalance present in the data for most of the defined scenarios, with the negative outcomes having significantly more observations than the positive outcomes. To deal with this problem, the authors decided to adapt the classification threshold for each scenario. The optimal threshold, that results in the highest AUC value, was calculated from the ROC curve, and set for each scenario.

For the KNN classifier, the authors chose the value of K by using a 5-fold CV and fitting multiple models with K varying between 1 and 25. The ROC was used to compare the models and to select the optimal K value for each scenario.

Fitting and hyperparameter tuning RF

The authors decided to use the packages *ranger* and *mlr* to both build and tune the models for the professional reseller. The authors are aware of the myriad of available methods and packages for this type of model, which all have advantages and disadvantages, but chose to opt for these mainly due to their increased efficiency

Since however the dataset is large and the methods are computationally expensive, the authors made sure their computers were using all cores during the fitting procedure, which allows for increased speed of computation.

At first, to build a baseline model, the authors opted for using the default parameters given by *ranger*. This led to models with $num.trees = 500$, $mtry = 4$ (the rounded down square root of the number of variables, see James et al. (2017)) and $min.node.size = 10$. This model was then used to predict the test set as a benchmark.

In the second step, the authors used the functionalities provided by the *mlr* package to tune the hyperparameters. In the default modelling, $mtry = 4$ was selected. This is explained by James et al. (2017), who describe that the splitting variables m, a random sample out of all the predictors available, is usually equal to the square root of the total of predictors available. Since the data in this case has 22 predictors, the square root equals 4.6. Hence, for tuning purposes, the authors compared how changing this to a range between 2 and 10 would impact the ACC and predictive power of the model. Additionally, the node size was included in the tuning, allowing for it to be a value between 5 and 15. The authors used a CV of 5-fold, in order to keep the computational time at a feasible amount. The number of trees was also kept at 500, also with the intention in mind to not increase the runtime too much.

The tuned models were then predicted on the test set and the resulting ACC compared to the untuned models. Although the tuned models often outperformed the untuned ones, the improvements were insignificant. However, in each case the model performing better was selected. All models can be found in the code documents.

Fitting and Hyperparameter Tuning XGBoost

In order to fit XGBoost models, the package *xgboost* was used. Tuning was again done with *mlr*.

As a first step in the fitting process for both the classification and regression models, the data needed to be prepared to fit the requirements of the *xgboost* package. The predictor variable was changed to a numeric “label”, while the data to both train and test the models needed to be a xgb.DMatrix.

For classification, logistic regression for binary classification was used for the objective function together with the metric AUC, which would allow the authors to later find the optimal threshold again. For the regression models, the *objective: “reg:squarederror”* was applied together with the metric of RMSE.

In a first, untuned version, the default parameters of the package were used. This was complemented with finding the optimal $nround$ using 5-fold CV. Having identified the best $nround$ value, this was used together with the default parameters in training the model on the train set. Each model was then predicted on the test set, and, in the case of classification problems, the optimal threshold for accuracy was identified.

Also for XGBoost, the authors attempted parameter tuning in order to improve accuracy and predictive power. Values were chosen in a range around the default values. A complete overview of the ranges of parameters used can be seen in the Table 5.4.

Table 5.4: XGBoost: Overview of the ranges of parameter tuning

	Lower	Upper
max_depth	3	10
min_child_weight	1	10
subsample	0.5	1
colsample_bytree	0.5	1

Again, the tuned models often outperformed the untuned ones. In each case, the model with the best ACC and smallest ERR was selected. All models can be found in the code documents.

Classification threshold

Two of the scenarios for the professional reseller contain classification models. Hence, the optimal threshold needed to be adjusted. Using the *cords()* function together with the ROC values per model (both tuned and untuned), the authors arrived at the optimal threshold and were able to predict each model according to the respective threshold.

In the following, the results of the models for each scenario are presented.

Results

Question 1: Should the reseller purchase a sneaker at retail price on the release date if they want to achieve a price premium of at least X% within 14 days?

Question 1 defined for the professional reseller revolves around the decision whether to purchase a sneaker at retail price on the release day if the aim is to resell it for a profit. In order to analyze this question, four scenarios were defined. Each scenario assumes a different level of price premium the reseller aims for. The observation period was limited to 14 days after release of a sneaker, as the professional reseller is characterized as aiming to sell as quickly as possible, at the highest profit possible. The aim was to predict the binary variable *buy* which is defined as 1 if the achievable price premium for a sneaker within 14 days after its release is greater than 0%/15%/25%/50%, and 0 otherwise.

Table 5.5 lists the AUC and ACC for each classification model and each scenario.

All in all, especially logistic regression, RF and XGBoost perform well on the train set and show high discriminatory power, which can be seen in their regularly high AUC values. The models are able to predict the positive “buy” outcome quite well, meaning they are all able to deal with the class imbalance in the dataset.

KNN only provides strong discriminatory power and high ACC in Scenario 1.1, while not reaching reliable high ACC values as the other three methods do for the remaining three scenarios.

It can be seen that on average, the complex models RF and XGBoost outperform the other two scenarios. This can be seen in both the higher AUC values and higher ACC values. This is especially evident for Scenario 1.2 and Scenario 1.3, where KNN shows the worst performance over all scenarios and XGBoost the best performance and highest ACC over all four alternatives.

Scenario 1.4 is a special case. Logistic regression shows the highest AUC, meaning the best ability to discriminate between the two classes, while its ACC is still slightly worse than that of RF. Despite its lengthier fitting process, XGBoost does not manage to predict the large positive price premium that well but predicts slightly more conservative, while still reaching a high AUC value.

Adding to this observation the fact that both RF and XGBoost went through a lengthier fitting and tuning process, and are more generally more complex models, one needs to consider the trade-off between higher ACC and more efficiency or rather speed when getting the results.

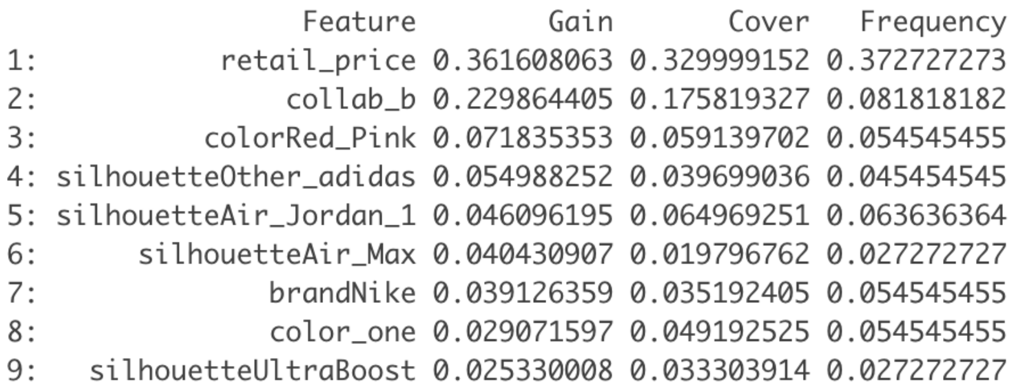
Table 5.5: Model results of Question 1 - AUC and ACC values

	Scenario 1.1		**Scenario 1.2**		**Scenario 1.3**		**Scenario 1.4**	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
Logistic Regression	0.746	70.63%	0.759	73.02%	0.786	74.60%	0.734	72.22%
KNN	0.711	81.75%	0.685	63.49%	0.645	57.14%	0.685	63.49%
RF	0.740	73.81%	0.799	73.81%	0.794	76.19%	0.717	74.60%
XGBoost	0.744	73.02%	0.788	74.60%	0.795	76.98%	0.728	63.49%

The professional reseller is described as mainly interested in most accurate results. This leads the authors to the conclusion that for this question, a mixture of both RF and XGBoost is most fitting to reliably predict whether or not the user should purchase the shoe. However, in order to find out which of the characteristics of a sneaker impact this decision the most, the coefficients of the logistic model are a helpful measure. Hence, the most important variables for XGBoost and the coefficients for logistic regression will be analyzed in the following.

Figure 5.1 illustratively shows the feature importance for XGBoost in Scenario 1.3. Since these are very similar in all scenarios, only one will be highlighted here. The measure "gain" is the measure classifying the features, giving information on the improvement in ACC brought by the respective feature to the branches it is on (xgboost developers, n.d.). It shows the relative contribution of the feature to the model overall, and hence a high value indicates high importance. It can be seen that the most important variable here is the feature *retail_price*, followed by *collab_b* and *colorRed_Pink*. This means that these variables increase the predictive power of the XGBoost model.

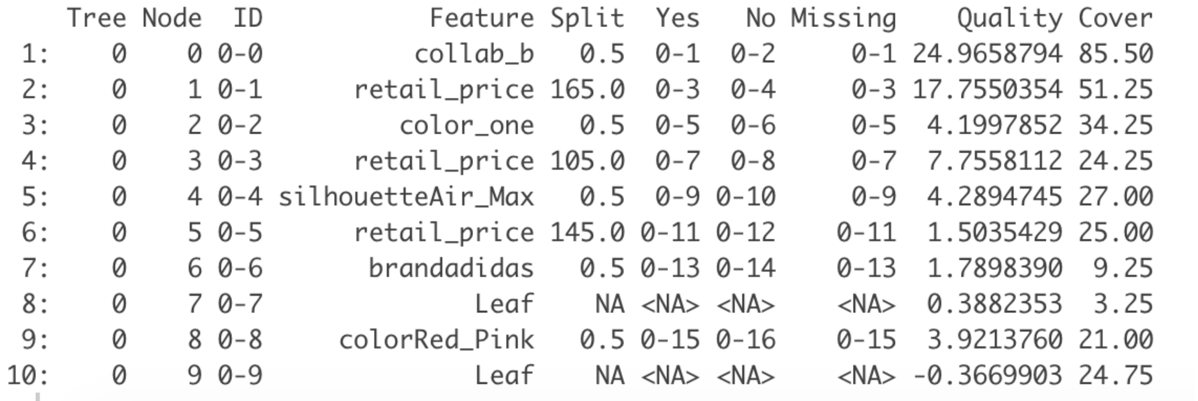
For all four scenarios and all four models, the most important variable is always the retail price. For XGBoost and RF, this is then followed by the indication *collab_b*, followed by various different colors and silhouettes as well as brands. It is not surprising that *retail_price* is so important for each of these models. Being one of the only non-binary variables in the dataset, this will clearly have significant impact on the ACC at each split.



```
                      Feature        Gain       Cover   Frequency
1:               retail_price 0.361608063 0.329999152 0.372727273
2:                   collab_b 0.229864405 0.175819327 0.081818182
3:              colorRed_Pink 0.071835353 0.059139702 0.054545455
4: silhouetteOther_adidas 0.054988252 0.039699036 0.045454545
5: silhouetteAir_Jordan_1 0.046096195 0.064969251 0.063636364
6:         silhouetteAir_Max 0.040430907 0.019796762 0.027272727
7:                  brandNike 0.039126359 0.035192405 0.054545455
8:                  color_one 0.029071597 0.049192525 0.054545455
9:     silhouetteUltraBoost 0.025330008 0.033303914 0.027272727
```

Figure 5.1: Feature importance for XGBoost in Scenario 1.3

Figure 5.2 shows a deeper dive into the important variables by providing information on the split measures. This tells the user the split that is applied to the respective feature on a branch of one of the trees. The split always indicates <, meaning in line 1, the split is performed at *collab_b* < 0.5. In the second line, the split is done at *retail_price* < $165 and third line, since this is again a binary variable, the split means counting those occurrences that are not *color_one*. The *yes* and *no* columns indicate the IDs of the following nodes, given the split condition before is met respectively not met. The IDs are noted in column 3, which allows the user to see that if the condition *collab_b* < 0.5 is met, the following node will be *retail_price* in line 2. The *quality* column gives insight into the split gain or, in the case of the line being a leaf, the value of this leaf. The plot overall shows how the first decisions are based on whether or not the retail price of a sneaker is released in collaboration and following this whether the retail price was below $165 or respectively the shoe is of *color_one*.

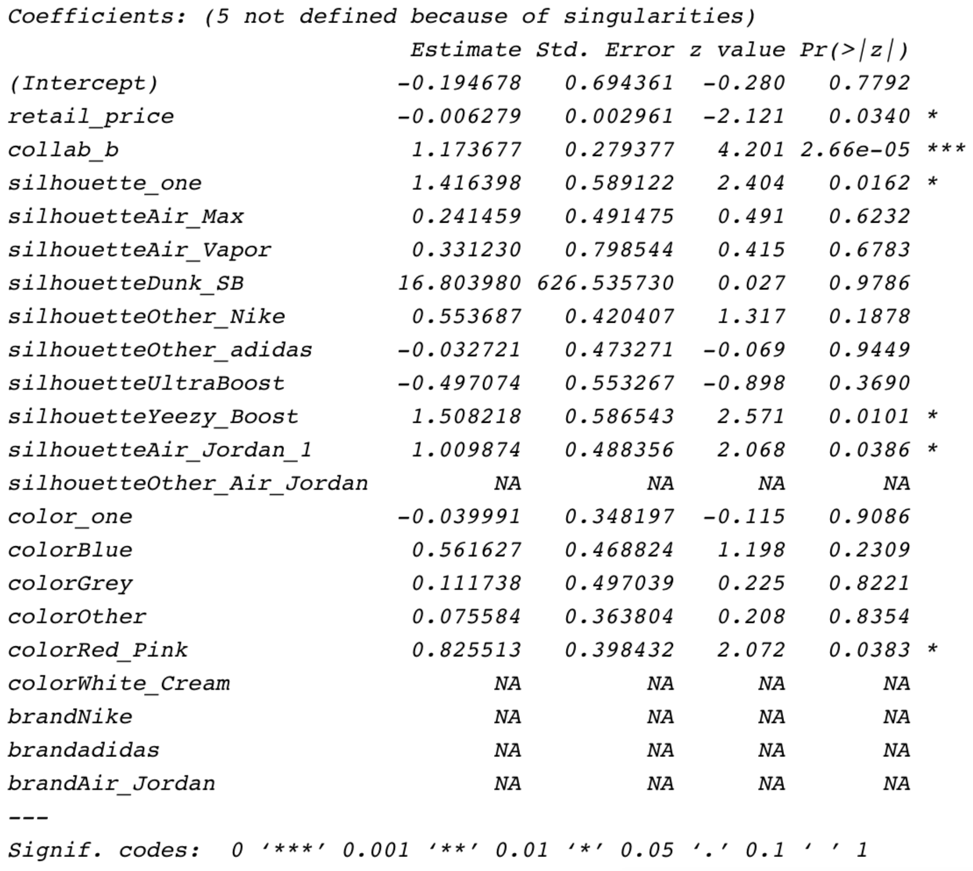


```
    Tree Node  ID           Feature Split  Yes   No Missing    Quality Cover
 1:    0    0 0-0          collab_b   0.5  0-1  0-2     0-1 24.9658794 85.50
 2:    0    1 0-1      retail_price 165.0  0-3  0-4     0-3 17.7550354 51.25
 3:    0    2 0-2         color_one   0.5  0-5  0-6     0-5  4.1997852 34.25
 4:    0    3 0-3      retail_price 105.0  0-7  0-8     0-7  7.7558112 24.25
 5:    0    4 0-4 silhouetteAir_Max   0.5  0-9 0-10     0-9  4.2894745 27.00
 6:    0    5 0-5      retail_price 145.0 0-11 0-12    0-11  1.5035429 25.00
 7:    0    6 0-6       brandadidas   0.5 0-13 0-14    0-13  1.7898390  9.25
 8:    0    7 0-7              Leaf    NA <NA> <NA>    <NA>  0.3882353  3.25
 9:    0    8 0-8     colorRed_Pink   0.5 0-15 0-16    0-15  3.9213760 21.00
10:    0    9 0-9              Leaf    NA <NA> <NA>    <NA> -0.3669903 24.75
```

Figure 5.2: Splitting criteria for XGBoost in Scenario 1.3

In order to gain more insight into how these variables impact the purchase decision, Figure 5.3 shows the coefficients for the logistic regression in Scenario 1.3. It was decided to include these as the logistic model does not perform substantially worse than the other models.

The most important variables are again *collab_b* and *retail_price*. Other important variables are e.g. *silhouetteYeezy_Boost* and *colorRed_Pink*. The coefficient of *retail_price* is negative, meaning that the larger the retail price of a sneaker, the smaller the probability that the model will predict *buy* = 1, meaning a smaller probability that the reseller will be able to reach the desired price premium within the time horizon of two weeks. This indicates that sneakers which are released at a smaller retail price are more likely to generate a higher price premium in the resale market shortly after release, than those shoes that were considered more expensive in retail. However, the variable *retail_price* is not as significant as e.g. *collab_b*. This variable has a positive value, meaning that a sneaker released in collaboration is more likely to increase the outcome *buy* to 1, meaning the reseller has more chance to reach his desired price premium. These are interesting first insights into how the resale price of a sneaker is impacted after release.



```
Coefficients: (5 not defined because of singularities)
                            Estimate Std. Error z value Pr(>|z|)    
(Intercept)                -0.194678   0.694361  -0.280   0.7792    
retail_price               -0.006279   0.002961  -2.121   0.0340 *  
collab_b                    1.173677   0.279377   4.201 2.66e-05 ***
silhouette_one              1.416398   0.589122   2.404   0.0162 *  
silhouetteAir_Max           0.241459   0.491475   0.491   0.6232    
silhouetteAir_Vapor         0.331230   0.798544   0.415   0.6783    
silhouetteDunk_SB          16.803980 626.535730   0.027   0.9786    
silhouetteOther_Nike        0.553687   0.420407   1.317   0.1878    
silhouetteOther_adidas     -0.032721   0.473271  -0.069   0.9449    
silhouetteUltraBoost       -0.497074   0.553267  -0.898   0.3690    
silhouetteYeezy_Boost       1.508218   0.586543   2.571   0.0101 *  
silhouetteAir_Jordan_1      1.009874   0.488356   2.068   0.0386 *  
silhouetteOther_Air_Jordan        NA         NA      NA       NA    
color_one                  -0.039991   0.348197  -0.115   0.9086    
colorBlue                   0.561627   0.468824   1.198   0.2309    
colorGrey                   0.111738   0.497039   0.225   0.8221    
colorOther                  0.075584   0.363804   0.208   0.8354    
colorRed_Pink               0.825513   0.398432   2.072   0.0383 *  
colorWhite_Cream                  NA         NA      NA       NA    
brandNike                         NA         NA      NA       NA    
brandadidas                       NA         NA      NA       NA    
brandAir_Jordan                   NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.3: Coefficient estimates of logistic model in Scenario 1.3

Question 2: How does the resale price of a sneaker develop in the first 14 days after its release?

Question 2 aims at addressing the professional reseller's interest in how the resale price of a sneaker varies in the short-term following its release. In order to analyze this question, the variable *price*, that is the current resale price of a sneaker, was predicted. Based on the authors' assumptions on the professional reseller (see Table 5.1), the observation period was limited to 14 days after a sneaker's release.

Table 5.6 lists the test MSE and the test RMSE for each model. Overall, none of the models show an excellent performance on this task. It can be seen clearly that the RF tuned for this model does not compare to the other three options, with an RMSE of more than double the one of XGBoost. Linear regression, KNN and XGBoost are somewhat better at predicting the price over this time frame, however XGBoost outperforms the former two. Comparing linear regression and KNN, although the test RMSE is better for linear regression, the adjusted R^2 of the linear model is lower than that of KNN (0.3192 vs. 0.4811).

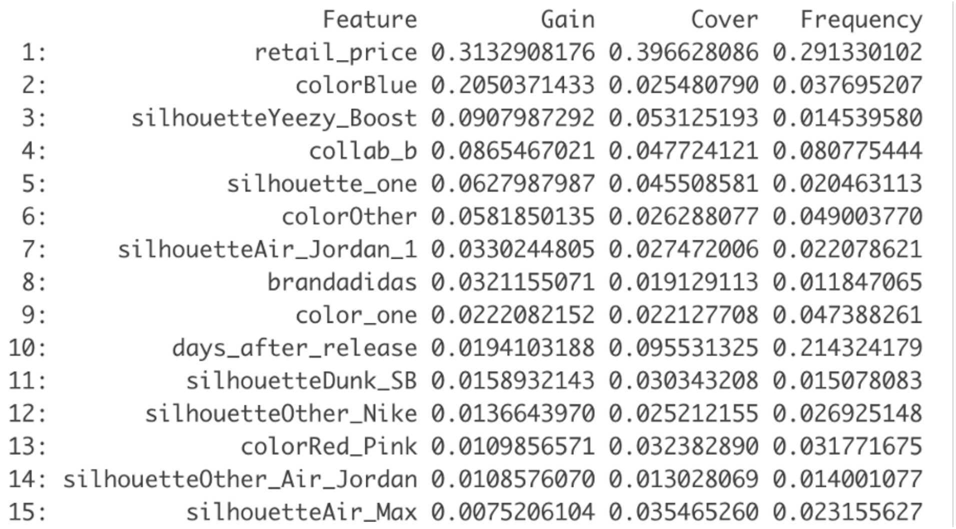
The authors assume that this is an example of where only the data is not sufficient in incorporating all factors influencing resale price and that additional influencing factors are not included in the dataset.

For a reseller to use these predictions, the XGBoost model provides the best alternative amongst the selection and is hence chosen for the further analysis.

Table 5.6: Model results for Question 2 – test MSE and test RMSE

	Test MSE	Test RMSE
Linear Regression	20575.21	143.44
KNN	15237.91	123.44
RF	79555.93	282.06
XGBoost	10605.68	102.98

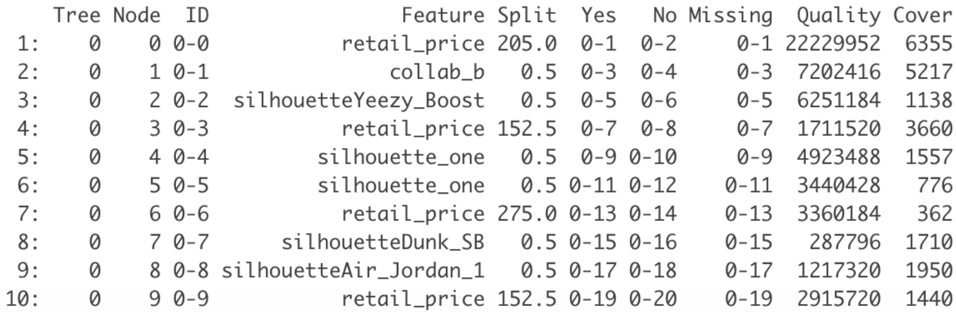
Figure 5.4 shows the most important variables for XGBoost in predicting the resale price. As previously, the most important variable for the prediction of the resale price is the initial retail price. This is followed by a mix of color and silhouette variables, showing how different characteristics of the sneakers influence the predictive power of the model in different ways. Seeing that some of the gain-values are very close to each other, this shows how the price prediction is driven quite equally by many different influencing variables.



```
                           Feature         Gain       Cover   Frequency
 1:                   retail_price 0.3132908176 0.396628086 0.291330102
 2:                      colorBlue 0.2050371433 0.025480790 0.037695207
 3:          silhouetteYeezy_Boost 0.0907987292 0.053125193 0.014539580
 4:                       collab_b 0.0865467021 0.047724121 0.080775444
 5:                 silhouette_one 0.0627987987 0.045508581 0.020463113
 6:                     colorOther 0.0581850135 0.026288077 0.049003770
 7:         silhouetteAir_Jordan_1 0.0330244805 0.027472006 0.022078621
 8:                    brandadidas 0.0321155071 0.019129113 0.011847065
 9:                      color_one 0.0222082152 0.022127708 0.047388261
10:              days_after_release 0.0194103188 0.095531325 0.214324179
11:               silhouetteDunk_SB 0.0158932143 0.030343208 0.015078083
12:            silhouetteOther_Nike 0.0136643970 0.025212155 0.026925148
13:                   colorRed_Pink 0.0109856571 0.032382890 0.031771675
14: silhouetteOther_Air_Jordan 0.0108576070 0.013028069 0.014001077
15:               silhouetteAir_Max 0.0075206104 0.035465260 0.023155627
```

Figure 5.4: Feature importance for XGBoost in Question 2

Figure 5.5 shows, as before, the important variables calculations for the XGBoost model. Again, the *split* column indicates the split that is applied to the respective feature on a branch of one of the trees. A feature can occur more than once, seeing e.g. the *retail_price* occurring in lines 1 and 4. The split always indicates <, meaning in line 1, the split is performed at *retail_price* < $205. In the second and third, since both contain binary variables, the split means counting those occurrences that are not *collab_b* or *silhouetteYeezy_Boost.* Again, this shows how the collaborations, and certain silhouettes are highly relevant for the determination of the price. It is interesting to see that *retail_price* is included several times, at different levels and with different values. In this model, clearly different silhouettes are most important in the predictions and hence relevant in arriving at the most accurate price predictions.



```
    Tree Node  ID                Feature Split  Yes   No Missing  Quality Cover
 1:    0    0 0-0           retail_price 205.0  0-1  0-2     0-1 22229952  6355
 2:    0    1 0-1               collab_b   0.5  0-3  0-4     0-3  7202416  5217
 3:    0    2 0-2  silhouetteYeezy_Boost   0.5  0-5  0-6     0-5  6251184  1138
 4:    0    3 0-3           retail_price 152.5  0-7  0-8     0-7  1711520  3660
 5:    0    4 0-4         silhouette_one   0.5  0-9 0-10     0-9  4923488  1557
 6:    0    5 0-5         silhouette_one   0.5 0-11 0-12    0-11  3440428   776
 7:    0    6 0-6           retail_price 275.0 0-13 0-14    0-13  3360184   362
 8:    0    7 0-7      silhouetteDunk_SB   0.5 0-15 0-16    0-15   287796  1710
 9:    0    8 0-8 silhouetteAir_Jordan_1   0.5 0-17 0-18    0-17  1217320  1950
10:    0    9 0-9           retail_price 152.5 0-19 0-20    0-19  2915720  1440
```

Figure 5.5: Splitting criteria for XGBoost in Question 2

Question 3: What is the maximum price premium that the reseller can achieve within 14 days after a sneaker's release?

The third question explores the professional reseller's interest in the maximum profit that can be realized with reselling a sneaker. In order to analyze this question, the variable *max_p_premium*, that is the maximum value for the price premium of a sneaker within 14 days after its release, is predicted. Again, the observation period is narrowed to 14 days after a sneaker's release time.

Table 5.7 lists the test MSE and the test RMSE for each model. The overview shows how bad both RF and XGBoost are able to reliably predict the price premium in comparison to linear regression and KNN. The RMSE reached here are 251.32 for RF and 219.32 for XGBoost, significantly higher than the values of linear regression (29.25) and KNN (99.38). This is surprising, since the general assumption, and also an assumption made by the authors, is that usually the more complex models also predict more accurately. However, in this case this cannot be applied, and the simplest models are also those with the highest predictability. Linear regression even significantly outperforms KNN, seeing the RMSE is less than 1/3 of that of KNN. Also the related adjusted R^2 value at 0.8867 is decent and gives additional argument for choosing this model in this case.

Table 5.7: Model results for Question 3 – test MSE and test RMSE

	Test MSE	**Test RMSE**
Linear Regression	855.75	29.25
KNN	9876.83	99.38
RF	63164.00	251.32
XGBoost	48099.90	217.50

Hence, the following will highlight and interpret the important variables and coefficients of the linear model for this scenario. Figure 5.6 shows the model's ten most important variables, which are based on the coefficients. The most important variables are the variables *price*, *retail_price* and *silhouetteDunk_SB*. At first glance, the authors find it highly interesting that the latter two are also included in the most important variables in the previous two models.

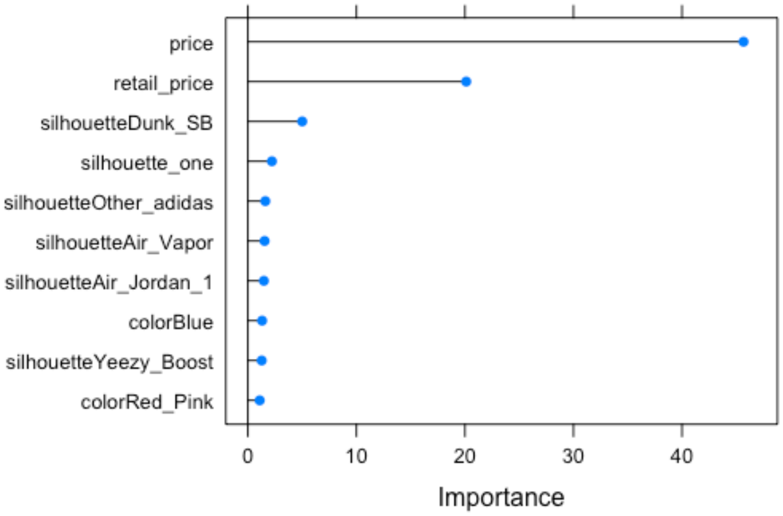
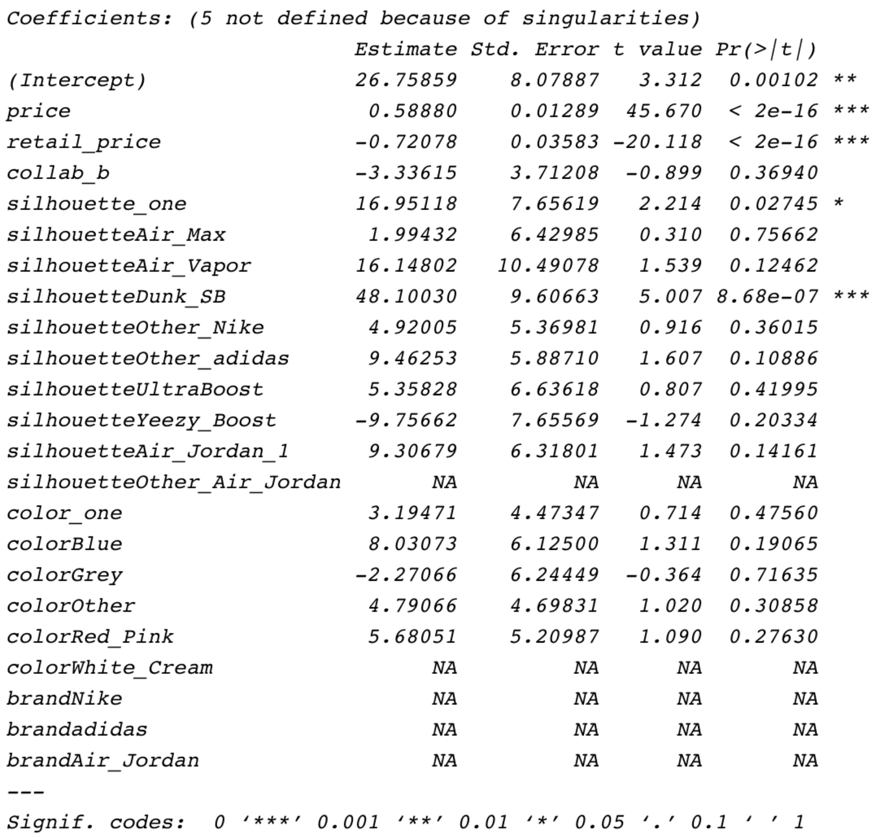


Figure 5.6: Variable importance plot of linear regression in Question 3

Although the silhouette is important, the value assigned to it is significantly lower than that of the retail price. Upon observing the coefficients, presented in Figure 5.7, it is highly interesting to see that the coefficient for *retail_price* is negative. This means that there is a negative correlation between retail price and the maximum price premium, which in turn means that the higher the retail price, the lower the maximum price premium the reseller can achieve in the first 14 days will be. The very positive coefficient of *silhouetteDunkSB* indicates that a shoe released with this silhouette will most likely increase in maximum price premium in the first 14 days. A reseller would hence be able to resell it at a large positive price premium. Furthermore, it is interesting to see that for this model, the top most important variables plots are dominated by silhouette-features, and other characteristics that seemed important in previous model building such as colors or the collaboration are not that significant in this case. Overall, it is clear that the leading indicators for a price premium prediction for the first 14 days are the retail price and the silhouette “DunkSB”.



```
Coefficients: (5 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               26.75859    8.07887   3.312  0.00102 ** 
price                      0.58880    0.01289  45.670  < 2e-16 ***
retail_price              -0.72078    0.03583 -20.118  < 2e-16 ***
collab_b                  -3.33615    3.71208  -0.899  0.36940    
silhouette_one            16.95118    7.65619   2.214  0.02745 *  
silhouetteAir_Max          1.99432    6.42985   0.310  0.75662    
silhouetteAir_Vapor       16.14802   10.49078   1.539  0.12462    
silhouetteDunk_SB         48.10030    9.60663   5.007 8.68e-07 ***
silhouetteOther_Nike       4.92005    5.36981   0.916  0.36015    
silhouetteOther_adidas     9.46253    5.88710   1.607  0.10886    
silhouetteUltraBoost       5.35828    6.63618   0.807  0.41995    
silhouetteYeezy_Boost     -9.75662    7.65569  -1.274  0.20334    
silhouetteAir_Jordan_1     9.30679    6.31801   1.473  0.14161    
silhouetteOther_Air_Jordan      NA         NA      NA       NA    
color_one                  3.19471    4.47347   0.714  0.47560    
colorBlue                  8.03073    6.12500   1.311  0.19065    
colorGrey                 -2.27066    6.24449  -0.364  0.71635    
colorOther                 4.79066    4.69831   1.020  0.30858    
colorRed_Pink              5.68051    5.20987   1.090  0.27630    
colorWhite_Cream                NA         NA      NA       NA    
brandNike                       NA         NA      NA       NA    
brandadidas                     NA         NA      NA       NA    
brandAir_Jordan                 NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.7: Coefficient estimates for linear regression in Question 3

Question 4: When should the reseller sell a sneaker if they aim to achieve a high price premium?

Question 4 explores the best time to resell that leads to a high price premium. In order to analyze this, three scenarios were defined. Each scenario assumes different days the reseller is willing to sell based on the price premium that can be achieved. Again, the observation period was limited to 14 days after release of a sneaker. The models aim to predict the binary variable *sell* which is defined as 1 for days where the price premium of a sneaker is highest/within the 3 highest values/within the 5 highest values, and 0 otherwise.

Table 5.8 lists the correct classification rate for each classification model and each scenario. In this case, logistic regression is outperformed in all three scenarios by the other three models. This can be seen both with the lower AUC ratios (below 0.7 in all scenarios) and lower ACC values (below 66% in all scenarios), meaning the model is not able to discriminate well between the two classes *sell* = 1 and *sell* = 0. In comparison, KNN, RF and XGBoost show AUC values of above 0.7 in all three cases as well as related high ACC scores of above 67%, showing that they have higher discriminatory power.

Interestingly, all models show the lowest ACC in Scenario 4.3, which is predicting the variable *sell* at the five highest price premiums. It could be assumed that a model predicting a larger target base (top five days rather than just top three days) would provide more accuracy, however this is not the case in the present example. The widest range of ACC can be seen for KNN: while in Scenario 4.1 it scores the highest percentage of 87%, it scores lowest out of these three models in Scenario 4.3 with an ACC of only 67%. It is again fascinating to see that in Scenario 4.1, KNN performs significantly better in terms of ACC than the other models. This was also the case for Question 1. Additionally insightful is the fact that nevertheless, the AUC values are not the best in comparison to RF and XGBoost. When comparing XGBoost and RF, which both went through a lengthy tuning process, it can be seen clearly that this is not justified for XGBoost which is in all three cases outperformed by RF. While the differences in both ACC and AUC are only insignificant in comparison to RF, the professional reseller aims for highest predictive power which in this case speaks for the usage of the RF and KNN models.

Table 5.8: Model results for Question 4 – AUC and ACC values

	Scenario 4.1		**Scenario 4.2**		**Scenario 4.3**	
	AUC	ACC	AUC	ACC	AUC	ACC
Logistic Regression	0.654	60.58%	0.681	63.64%	0.671	65.53%
KNN	0.750	87.91%	0.761	77.71%	0.709	67.61%
RF	0.799	77.71%	0.790	72.57%	0.733	69.97%
XGBoost	0.791	76.68%	0.754	72.33%	0.715	68.74%

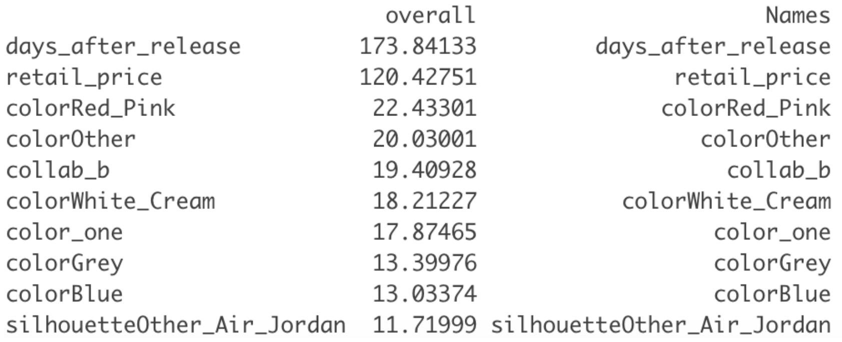
Although KNN is outperforming the other models in terms of accuracy in Scenarios 4.1 and Scenario 4.2, which is the main aim of the professional reseller, the model does not estimate coefficients and does not provide additional insight into the importance of variables. Therefore, the authors decided to look at the important variables suggested by the RF models in order to gain more insight into the predictions in this case.

In all three scenarios, the two most important variables in the RF models are the same: (1) *days_after_release* and (2) *retail_price* (see Figure 5.8 for example). This means that these two variables provide the largest improvement of the predictive power of the models. This indicates that the timing of when a shoe is traded in relation to the release day impacts the outcome variable *sell* significantly. Additionally, the retail price has an impact on the outcome.

It is unfortunately not possible to analyze in which way this impacts the outcome. However, both variables can also be seen as important as they are non-binary, while the majority of remaining variables are binary, which could also lead to increased importance measures. These findings however seem to give indication that timing as well as the initial price impact price premium prediction, and therefore resale price prediction, significantly.

However, it additionally pays to take a look at the remaining variables. It can be seen that there is a large difference between the values of the first two variables and the ones following, while the latter variables score values close to each other.

Across all three scenarios, it is interesting to see that color plays an important role. As an example, Figure 5.8 shows the feature importance for Scenario 4.1, where one can see that apart from the final variable *silhouetteOther_Air_Jordan* as well as the indicator for collaboration, all other following variables are color variables. Solely from this model and the overview beneath, one can see that the prediction of when to resell when aiming for a high price premium is significantly impacted by the color of a shoe and not majorly, as could be seen before, by one specific silhouette and a mixture of other characteristics.



```
                              overall                      Names
days_after_release          173.84133         days_after_release
retail_price                120.42751               retail_price
colorRed_Pink                22.43301              colorRed_Pink
colorOther                   20.03001                 colorOther
collab_b                     19.40928                   collab_b
colorWhite_Cream             18.21227           colorWhite_Cream
color_one                    17.87465                  color_one
colorGrey                    13.39976                  colorGrey
colorBlue                    13.03374                  colorBlue
silhouetteOther_Air_Jordan   11.71999 silhouetteOther_Air_Jordan
```

Figure 5.8: Feature importance for RF in Scenario 4.1

5.3.2 Mixed-Role Reseller

The mixed-role reseller is assumed to be less interested in maximizing price or price premium, since the main aim is the collection of these sneakers. The authors hence assume a lower need for high accuracy of models, and more interest in interpretability. To provide for these needs and the respective scenarios for this persona, the authors used the same models as before: linear and logistic regression, KNN, RF and XGBoost. Additionally, to provide a way of interpretability, the authors added a simple decision tree model for a mixed-role reseller to be able to easily understand the predictions. The modelling process is described in the next section. This is followed by the analysis of the results.

Modelling

Fitting Linear Regression, Logistic Regression and KNN

For the scenarios of the mixed-role reseller, the authors followed the same approach in fitting the models from linear regression, logistic regression and KNN as for the professional reseller. The optimal classification thresholds for the logistic models were calculated from the ROC curves for each scenario. Again, for the KNN classification multiple models with varying K values were built and 5-fold CVs were performed to select the optimal value for K in each scenario.

Fitting of Regression Tree

Due to its easy interpretability, the simple regression tree was included in the determination of best models for the mixed role reseller. The authors aimed to make use of the potential to plot the results, to highlight to non-machine learning users how the algorithm arrives at the decision making. The trees were fitted on the train data set using the *rpart* package, which allows the adding of a *control* element with a 5-fold CV. Additionally, using the package *rpart.plot*, the results were displayed in an easily interpretable way.

Fitting of RF

Since one of the main assumptions for the mixed-role reseller is the need for simplicity of the models, the authors decided to not tune the RF models for this persona.

For the majority of scenarios, the package used for building the RF model was again *ranger*. On one occasion, due to significantly better performance, the authors chose a model built using the *caret* package. Again, the authors made sure their computers were using all cores during the fitting procedure.

The RF model built was the model using default parameters available in *ranger,* that is, $num.tree = 500, mtry = 4,$ and $min.node.size = 10$. In the case of using *caret,* the model was build using the *method = "rf"*.

The nature of RF models, in which a random sample of predictors is taken for each new tree, already reduces the danger of overfitting. Furthermore, RF models built with *ranger* calculate automatically the Out Of Bag-score, which means a part of the data will be excluded from the fitting process and the model will be tested on this data during the fitting already. Hence, to save time and keep the modelling simple, the authors decided to remove an additional CV procedure in this case. For the model built using *caret*, the 5-fold CV was added.

The models were then predicted on the test set using the same mechanisms as described before on reaching the best threshold for classification using the *cords()* function for ROC and analyzing error rates such as RMSE.

Fitting of XGBoost

Although XGBoost does not offer easy interpretability, the authors decided to again include the base model in this part of the analysis. Again, no tuning was performed, which decreased the fitting time significantly and better fits the assumptions made in relation to the mixed-role reseller. The potential loss of accuracy was taken as a trade-off to be explained with the aim of the mixed-role reseller being not the highest accuracy, but rather feasibility and time. As before, the package *xgboost* was used for the process.

The data preprocessing and fitting of the base model was performed accordingly to the fitting of the based model for the professional reseller (see Chapter 5.3.1) and according to the scenario requirements (classification or regression).

As before, the models were predicted on the test set using the same mechanisms as described before.

Results

Question 5: Should the reseller purchase a sneaker at the current resale price if they want to achieve a price premium of at least X% on day 180 after release?

Question 5 revolves around the mixed-role reseller's decision whether to buy a sneaker if the aim is to resell after a longer holding period while keeping the costs incurred low. In order to analyze this question, three scenarios were defined. Each scenario assumes a different level of price premium the reseller aims for. As the mixed-role reseller's primary motivation is collecting and not maximizing profits, they are willing to buy sneakers that bring lower price premium levels, or even a negative price premium. The mixed-role reseller's price premium was calculated based on the resale price on the day of purchase and the resale price on day 180 (see Table 5.1). The observation period was limited to 30 days after release, as the mixed-role reseller is characterized as aiming to purchase a sneaker within the first 30 days after release.

The aim was to predict the binary variable *buy* which is defined as 1 if the achievable price premium for a sneaker on day 180 after its release is greater than 5%/0%/-5%, and 0 otherwise.

Table 5.9 lists the resulting AUC and ACC for each classification method and each scenario. With regards to the AUC values, the XGB classifier outperforms in all scenarios, meaning that the models are able to distinguish very well between the two classes. Further, the resulting predictions have a relatively high ACC ranging between 76% and 78% for the different scenarios. Also, the models built with RF and KNN have a high discriminative power and perform strongly, with ACC values ranging from 75% to 78%, and 71% to 77%, respectively. The remaining two methods, logistic regression and decision trees, result in less accurate predictions and the in comparison lower AUC values indicate lower discriminative power. Although the more complex models, RF and XGB, are more accurate, the KNN classification could be the best alternative for the mixed-role reseller, who is more interested in simplicity in a method, than accuracy. Models with KNN are simple and easy to fit, and ACC rates that are acceptable for the mixed-role reseller are achievable in these scenarios.

Table 5.9: Model results for Question 5 - AUC and ACC values

	Scenario 5.1		**Scenario 5.2**		**Scenario 5.3**	
	AUC	ACC	AUC	ACC	AUC	ACC
Logistic Regression	0.682	65.80%	0.685	64.22%	0.676	62.88%
KNN	0.809	77.29%	0.809	73.73%	0.802	71.50%
Decision Tree	0. 537	69.62%	0.661	68.75%	0.666	63.85%
RF	0.845	78.59%	0.840	75.61%	0.831	75.17%
XGBoost	0.906	76.61%	0.905	78.75%	0.895	77.70%

KNN does not estimate coefficients that could suggest the importance of variables, hence, the authors decided to look at the variable importance plots and coefficients of the logistic models in order to create insights on how the mixed-roles reseller's purchase decision is affected by different variables. As the important variables do not vary between the three scenarios, only the variable importance plot for Scenario 5.1 is presented here. Figure 5.9 shows the ten most important variables based on the model's coefficients. The three most important variables are (1*) silhouetteUltraboost*, (2) *silhouetteAirMax*, and (3) *retail_price.* Exploring the coefficient estimates in Figure 5.10 can give some more insights. The coefficient estimates for the two most important variables*, silhouetteUltraboost* and *silhouetteAirMax*, are both negative. This indicates that if a sneaker's silhouette is not of the model "Ultraboost" or "Air Max", the mixed-role reseller is less likely to achieve his desired price premium in later resale, and hence, is less likely to purchase the sneaker. The coefficient estimate for *retail_price* is also negative, so the higher the initial retail price of a sneaker, the less likely is the reseller to achieve his expected price premium through resale and less likely to purchase the sneaker.

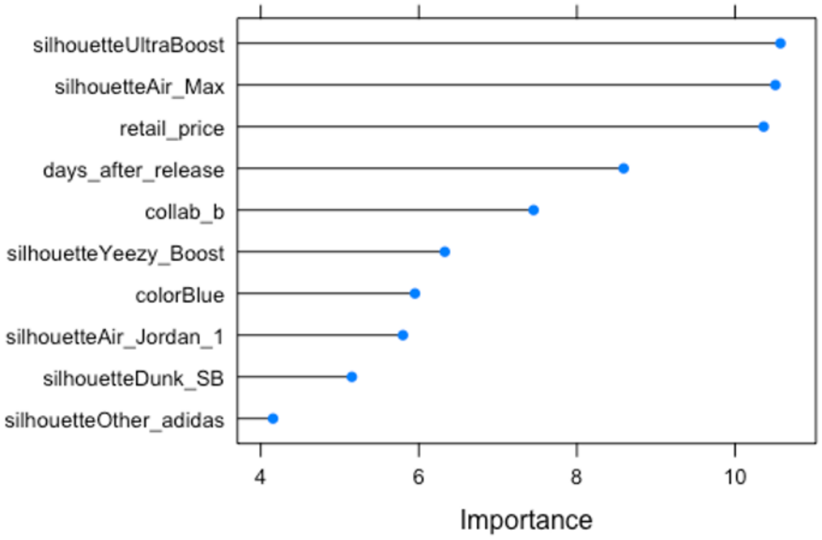
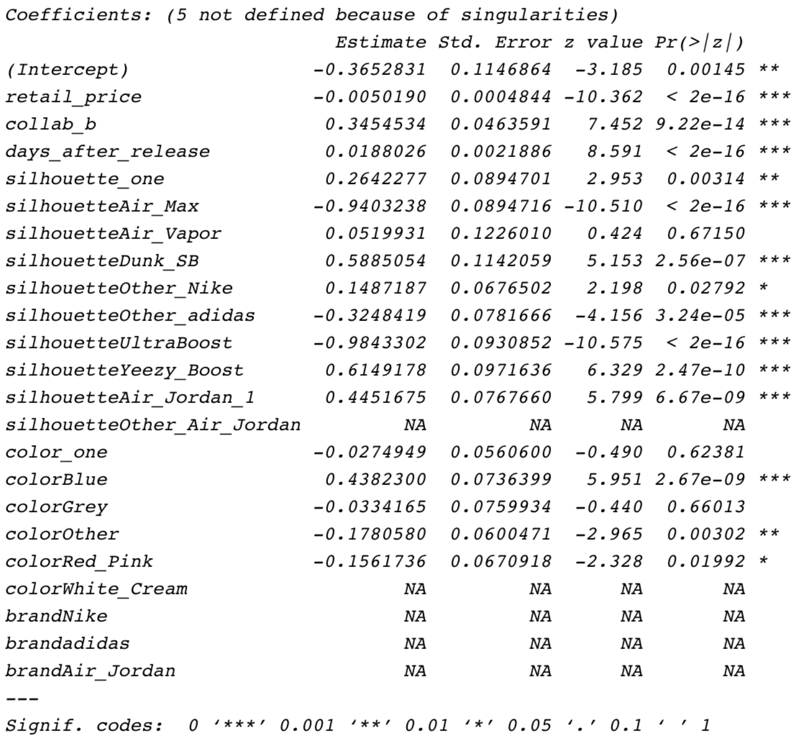


Figure 5.9: Variable importance plot of logistic regression in Scenario 5.1

The positive coefficient estimates for *collab_b* and *days_after_release*, which are also important variables in the model, reveal further interesting insights. The positive value of the coefficient *collab_b* suggests that if a sneaker is released as part of a collaboration, the reseller is more likely to achieve his expected price premium through resale, and hence, more likely to make a positive purchase decision. Further, the positive coefficient estimate for *days_after_release* indicates that with more days passing after a sneaker's release, the more likely the reseller is to generate his expected price premium through resale, and the more likely to buy the sneaker. This could hint that the resale value of a sneaker decreases with the days after its release, as the mixed-role reseller is more likely to achieve his price premium that is calculated based on the current resale price.



```
Coefficients: (5 not defined because of singularities)
                             Estimate Std. Error z value Pr(>|z|)    
(Intercept)                -0.3652831  0.1146864  -3.185  0.00145 ** 
retail_price               -0.0050190  0.0004844 -10.362  < 2e-16 ***
collab_b                    0.3454534  0.0463591   7.452 9.22e-14 ***
days_after_release          0.0188026  0.0021886   8.591  < 2e-16 ***
silhouette_one              0.2642277  0.0894701   2.953  0.00314 ** 
silhouetteAir_Max          -0.9403238  0.0894716 -10.510  < 2e-16 ***
silhouetteAir_Vapor         0.0519931  0.1226010   0.424  0.67150    
silhouetteDunk_SB           0.5885054  0.1142059   5.153 2.56e-07 ***
silhouetteOther_Nike        0.1487187  0.0676502   2.198  0.02792 *  
silhouetteOther_adidas     -0.3248419  0.0781666  -4.156 3.24e-05 ***
silhouetteUltraBoost       -0.9843302  0.0930852 -10.575  < 2e-16 ***
silhouetteYeezy_Boost       0.6149178  0.0971636   6.329 2.47e-10 ***
silhouetteAir_Jordan_1      0.4451675  0.0767660   5.799 6.67e-09 ***
silhouetteOther_Air_Jordan         NA         NA      NA       NA    
color_one                  -0.0274949  0.0560600  -0.490  0.62381    
colorBlue                   0.4382300  0.0736399   5.951 2.67e-09 ***
colorGrey                  -0.0334165  0.0759934  -0.440  0.66013    
colorOther                 -0.1780580  0.0600471  -2.965  0.00302 ** 
colorRed_Pink              -0.1561736  0.0670918  -2.328  0.01992 *  
colorWhite_Cream                   NA         NA      NA       NA    
brandNike                          NA         NA      NA       NA    
brandadidas                        NA         NA      NA       NA    
brandAir_Jordan                    NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.10: Coefficient estimates for logistic regression in Scenario 5.1

Question 6: How does the average monthly resale price of a sneaker develop in the first 6 months after its release?

Question 6 aims to address the mixed-role reseller’s interest in how the monthly resale price of a sneaker varies in the longer term. In order to analyze this question, the variable *av_p_month,* that is the average monthly resale price of a sneaker, was predicted.

Table 5.10 lists the test MSE and test RMSE for each fitted model. In comparison, the linear regression model and KNN model result in the highest error measure, suggesting that these simple models do not fit the data set well. Further, the linear model has an adjusted R^2 value of only 0.0791, so not much of the variability in the average monthly price is explained by the model’s independent variables.

Table 5.10: Model results for Question 6 - test MSE and test RMSE

	Test MSE	**Test RMSE**
Linear Regression	42.80	6.54
KNN	45.83	6.77
Regression Tree	0.9327	0.9657
RF	1.726E-04	0.01314
XGBoost	3.30E-09	5.74E-05

The best results are achieved with the model built with XGBoost, followed by the model built with RF. However, these two models might be too complex for the mixed-role-reseller who is primarily interested in simplicity in a model. So, the model constructed from regression trees could be the best alternative for the mixed-role-reseller, as it results in a fairly low test RMSE, and as it provides a visual output that is easy to interpret. Figure 5.11 shows a plot of the constructed tree. Based on the reduction of the MSE, the predictor variable *month*, that is the month in a year, was identified as the most important variable. The *month* variable seems to be dominant, as the resulting tree makes multiple splits of the same predictor. Still, it is a good example to demonstrate the easy interpretability of regression trees. For instance, the reseller can easily read from the tree that a sneaker's average monthly resale price is predicted to be $185 for the months March, April and May (see branch on the far left).

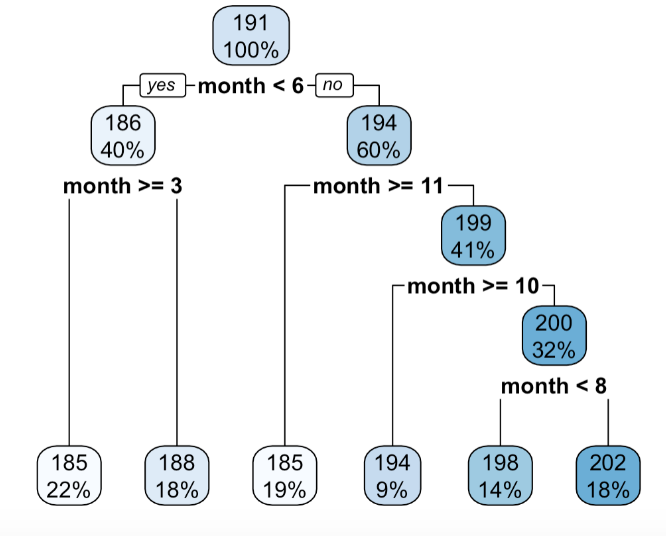


Figure 5.11: Constructed regression tree for Question 6

Question 7: When should the reseller buy a sneaker if they aim to pay a low price premium for it?

Question 7 explores the best purchase time that leads to the mixed-role reseller paying a low price premium for a sneaker on the secondary market. To analyze this question, three scenarios were defined. Each scenario assumes different days the reseller is willing to buy based on the price premium (compared to the retail price) they have to pay. The mixed-role-reseller pays the current resale price and purchases a sneaker within the first 30 days after its release (see Table 5.1). The models aimed to predict the binary variable *buy* which is defined as 1 for days where the price premium of a sneaker is lowest/within the 3 lowest values/within the 5 lowest values, and 0 otherwise.

The AUC and ACC values for each classification model and each scenario are listed in Table 5.11. While the XGBoost models slightly outperform in two out of three scenarios, still, both the models from RF and XGBoost perform relatively well in terms of AUC and overall ACC.

On the other hand, the models built with KNN and regression tree fail to deal with the class-imbalance present in the data set in Scenario 7.1 and Scenario 7.2 (and Scenario 7.3 for regression tree, marked as yellow in Table 5.11), hence, the models predict a negative outcome, i.e. *buy=0*, for most of the observations. As the mixed-role reseller has a strong interest in purchasing a sneaker for their collection, these models, predicting only negative outcomes, do not create any value for the reseller during the purchase decision process. For Scenario 7.3, where the data set is more balanced, the KNN model also predicts positive outcomes and achieves a relatively high ACC rate.

Models from logistic regression can better deal with the class-imbalance in the data set, that is, the models do not only predict a negative outcome. Compared to the more complex models RF and XGBoost, logistic regression models result in lower AUC values, and thus, have lower discriminatory power in comparison. However, the classification accuracy, except in Scenario 7.1, does not differ much compared to the more complex models from RF and XGBoost. Hence, the simpler models built with logistic regression may be the best alternative for the mixed-role reseller, who is presumed to prioritize simplicity of a model over accurate predictions.

Table 5.11: Model results for Question 7 - AUC and ACC values

	Scenario 7.1		**Scenario 7.2**		**Scenario 7.3**	
	AUC	ACC	AUC	ACC	AUC	ACC
Logistic Regression	0.708	64.27%	0.681	64.75%	0.675	62.94%
KNN	0.733	93.43%	0.705	85.97%	0.694	78.63%
Regression Tree	0.500	94.00%	0.500	86.02%	0.500	79.60%
RF	0.755	74.16%	0.728	65.31%	0.725	66.04%
XGBoost	0.752	79.27%	0.730	66.46%	0.727	66.27%

Looking at the importance of the model's variables can provide some insights into how the purchase decision is affected by various variables. The important variables in the logistic regression models do not vary much across the three scenarios, so exemplary, the variable importance plot for Scenario 7.1 is presented in Figure 5.12, which shows the ten most important variables. The importance of each variable is calculated based on the coefficients of the logistic model, which are displayed in Figure 5.13. In this case, the three most important variables are (1) *days_after_release*, (2) *collab_b,* and (3) *retail_price*.

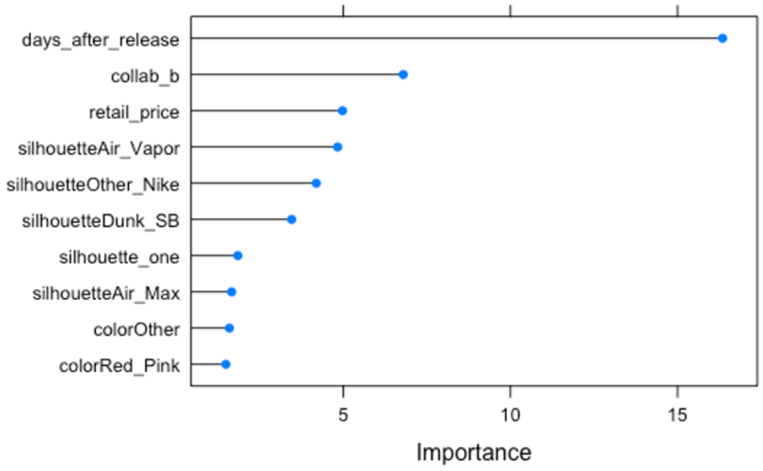
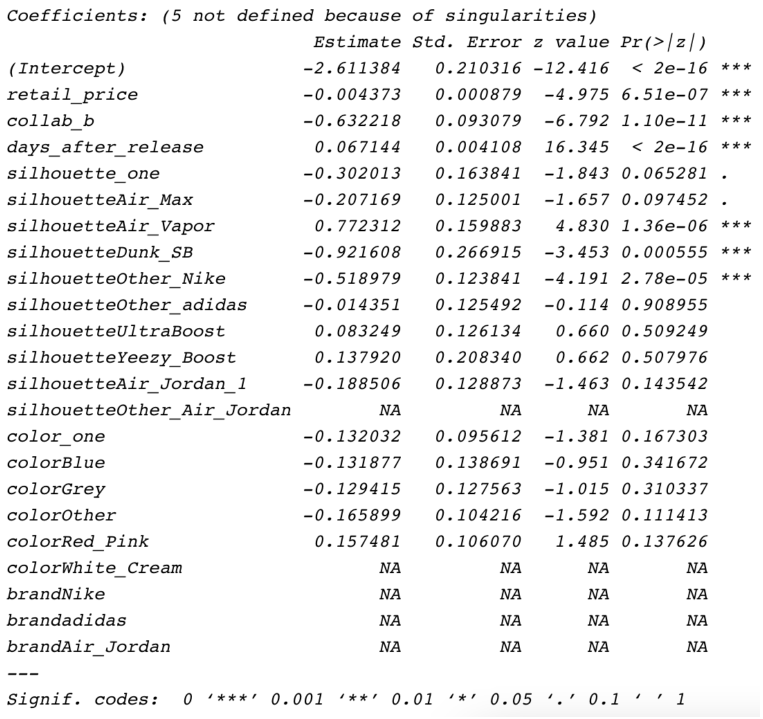


Figure 5.12: Variable importance plot of logistic regression in Scenario 7.1

The coefficient estimate of the most important variable *days_after_release* is positive, meaning that the more days after a sneaker's release pass, the more likely it is that the reseller purchases a sneaker by paying a low price premium. Finally, the coefficient estimate for *retail_price* is negative, so the higher the initial retail price of a sneaker, the less likely the reseller is to purchase the sneaker by paying a low price premium.



```
Coefficients: (5 not defined because of singularities)
                           Estimate Std. Error z value Pr(>|z|)    
(Intercept)               -2.611384   0.210316 -12.416  < 2e-16 ***
retail_price              -0.004373   0.000879  -4.975 6.51e-07 ***
collab_b                  -0.632218   0.093079  -6.792 1.10e-11 ***
days_after_release         0.067144   0.004108  16.345  < 2e-16 ***
silhouette_one            -0.302013   0.163841  -1.843 0.065281 .  
silhouetteAir_Max         -0.207169   0.125001  -1.657 0.097452 .  
silhouetteAir_Vapor        0.772312   0.159883   4.830 1.36e-06 ***
silhouetteDunk_SB         -0.921608   0.266915  -3.453 0.000555 ***
silhouetteOther_Nike      -0.518979   0.123841  -4.191 2.78e-05 ***
silhouetteOther_adidas    -0.014351   0.125492  -0.114 0.908955    
silhouetteUltraBoost       0.083249   0.126134   0.660 0.509249    
silhouetteYeezy_Boost      0.137920   0.208340   0.662 0.507976    
silhouetteAir_Jordan_1    -0.188506   0.128873  -1.463 0.143542    
silhouetteOther_Air_Jordan       NA         NA      NA       NA    
color_one                 -0.132032   0.095612  -1.381 0.167303    
colorBlue                 -0.131877   0.138691  -0.951 0.341672    
colorGrey                 -0.129415   0.127563  -1.015 0.310337    
colorOther                -0.165899   0.104216  -1.592 0.111413    
colorRed_Pink              0.157481   0.106070   1.485 0.137626    
colorWhite_Cream                 NA         NA      NA       NA    
brandNike                        NA         NA      NA       NA    
brandadidas                      NA         NA      NA       NA    
brandAir_Jordan                  NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5.13: Coefficient estimates for logistic regression in Scenario 7.1

5.3.3 Summary

The previous analysis included the formulation of two reseller personas, including their aims, needs and questions they could be interested in. This was followed by a presentation of modelling results and discussion of how the models would provide decision support for the previously formulated reseller questions bearing in mind different scenarios and target variables. The analysis showed how the different reseller needs and characteristics result in varying decisions and how these can be explored in different ways with different methods. The analysis further showed how the particular machine learning models provided varying results, sometimes similar results, sometimes results with large disparities.

Table 5.12 gives a short summary of the main findings from the analysis, including the most accurate methods and most important variables for each explored question. Further, the table summarizes the best identified model for the reseller persona based on their needs. For the professional reseller, who is primarily interested in model accuracy, a variety of methods can provide decision support for the previously formulated reseller questions. Models built with XGBoost often provided accurate predictions for the professional reseller, however, also simple models like linear regression and KNN performed best in some cases.

On the other hand, XGBoost models resulted in the most accurate predictions in all cases of the mixed-role reseller. However, the mixed-role reseller is assumed to prefer simplicity of a model over accuracy. The analysis showed that the simpler models built with KNN, regression trees, and logistics regression, result in relatively accurate predictions, and hence, were identified as suitable alternatives for the mixed-role-reseller.

Table 5.12: Summary of the analysis

Question	**Reseller Persona**	**Most accurate methods**	**Best alternative for persona**	**Most important variables**
Q1	Professional	RF with ACC values 74%-76% and XGBoost with ACC values 73%-77%	RF and XGBoost	retail_price collab_b silhouetteOther_Nike
Q2	Professional	XGBoost with test RMSE 102.98	XGBoost	retail_price colorBlue silhouetteYeezy_Boost
Q3	Professional	Linear Regression with test RMSE 29.25	Linear Regression	price retail_price silhouetteDunkSB
Q4	Professional	KNN with ACC values 67%-89%	KNN	days_after_release retail_price collab_b
Q5	Mixed-Role	XGBoost with ACC values 77%-79%	Simpler method KNN with ACC 72%-77%	silhouetteUltraBoost silhouetteAirMax retail_price
Q6	Mixed-Role	XGBoost with test RMSE 5.74E-05	Interpretable regression tree with test RMSE 0.9657	month
Q7	Mixed-Role	XGBoost with ACC values 60%-79%	Simpler method logistic regression with ACC values 63%-65%	days_after_release collab_b retail_price

6. Discussion

The following paragraph will discuss the hypotheses formulated in Chapter 2.4 and will conclude on whether the authors accept or reject these.

H1: The characteristics of a shoe have different impact on the price development.

Hypothesis 1 states that the characteristics of a shoe have different impact on the price development. The authors arrived at this hypothesis on the basis of the findings from the hedonic pricing model, stating that the price of an object is made up from the sum of the values of its characteristics. The authors were interested to see whether it would be possible to make investment decisions based on selected features of a sneaker. This would be especially important to know for the professional reseller.

The findings from the analysis in Chapter 5.3 give a basis for argumentation into two directions, regarding modelling results and business implications.

Regarding solely the results of the modelling procedures, one can clearly see how for each scenario and model used, different variables are important. This indicates that the variables used in predicting the various scenarios change, and hence information on different characteristics of a shoe and resale time improve the precision of the respective models. As an example, in Questions 4, which relates to the days when a reseller should sell his stock in order to gain a high price premium, the number of days after release are highly important. Additionally, in order to determine the best day, the color of a shoe plays a critical role, which can be seen in Figure 5.8, showing the important variables of RF in this case.

Furthermore, taking e.g. the important variable results from XGBoost in Question 2 as an example, one can clearly see how for example the silhouette “Dunk SB”, as well as the question whether a shoe was made in collaboration with someone, are highly relevant in improving the predictive power of the model and therewith, in this case, the determination of the resale price of a sneaker.

In order to give insight into the actual decision of purchasing or selling, it pays to review the coefficients which are provided by the linear and logistic models. As an example, the coefficients provided for Question 5 give indication into key characteristics to look out for if an investor wants to receive a certain price premium after 6 months. It can be seen that e.g.

the variable *silhouetteUltraboost*, one of the critical variables for this model, has a negative coefficient estimate. This means that a sneaker with this silhouette will not allow the reseller to achieve the desired price premium and hence would speak against him buying this sneaker in the first place. The positive coefficient estimate of collaboration in this case however shows that a sneaker released in collaboration will be more likely to increase in price and provide the desired price premium. Hence, the availability of coefficient estimates for the important variables of linear and logistic models can support investment decisions as they indicate how the outcome, may this be the price, price premium or solely a binary variable, will be impacted by this characteristic.

The authors hence accept the first hypothesis and conclude that the characteristics of a sneaker have different impact on its resale price development.

H2: The price increase on the resale market is related to the brand that released the sneaker in the first place.

The findings relate very much to the findings based on which Hypothesis 2 will be discussed. This states that the price increase on the resale market is related to the brand that released the sneaker in the first place. When formulating this hypothesis, the authors assumed that since the brand played and continues to play a significant role in the sneakerhead history and culture, it would be of highest importance in the decision-making processes. The authors were hence surprised to see that this is not always the case. Basing this finding on the important variables as well as coefficients again, as above for Hypothesis 1, it was clearly to see that the brand was hardly mentioned as an important variable.

However, seeing that specific silhouettes are frequently in the top important variables, this indicates that the brand is nevertheless indirectly relevant, but not the main criteria. The authors conclude that the resale price increase is, at most, indirectly related to the brand releasing the sneaker in the first place and more directly impacted by characteristics such as the silhouette. This shows a further interesting finding: it seems that after all the shoes might be more important than just the brand. The authors can hence neither accept nor reject the hypothesis entirely and suggest that this offers room for further research.

H3: The resale value of a sneaker is at peak directly after release and decreases then. Hypothesis 3 states that the resale value of a sneaker is at peak directly after release and decreases then. The authors came to this assumption based on their presumption that hype around a release builds up before the release date, climaxes on the day of release and then decreases again.

Again, in order to discuss this hypothesis, the authors refer to the results from the analysis. It can be seen for example in Question 7, defining when a mixed-role reseller should buy a sneaker if the aim is a low price premium, the coefficient estimate of the important variable days after release is positive. This means that the more days after a sneaker's release have passed, the higher the probability of the reseller purchasing it at a low, or even lowest, price premium. This indicates that the more days after release, the lower the minimum price premium. It however does not give insight into the highest price premium to be expected, and hence cannot be used as a full conclusion saying that the price will always decrease.

Further insight can be gained e.g. from the results related to Question 2, which predicts the price development of the sneakers 14 days after release. Although the linear model is not the best model for this matter, the coefficients show that the estimated coefficient for the variable days after release is negative, again indicating price decreases when more days pass.

However, contradicting indications can be seen in the results from Question 4 and Question 1. Question 4 dealing with the optimal selling date in order to gain a high price premium indicates as well that the amount of time after release impacts this decision. This means that in finding the day within the first 14 days after release on which the price premium is at top levels is impacted by the time passed. However, the models delt well with the class imbalance, did not only predict 0 (do not sell), but also predicted sell at times, which means that indeed the premium was high on some days. However, the response variable here was to sell if the price premium is at maximum, top 3 or within top 5 – without giving clear indication if these are positive or negative premiums. Hence, it can be the case that the highest price premiums are positive or negative.

This is however resolved with Question 1, identifying whether a reseller can achieve a certain (positive) price premium within 14 days after release. In these models, the days after release variable is not as important as e.g. the silhouette. However, seeing that again the models delt well with the class imbalance, had high AUC values and high prediction accuracy, they accurately predicted *buy* several times. This indicates that the price premium within the first 14 days is at times positive.

Hence, on the basis of the findings in this research, the hypothesis can neither be accepted nor declined completely. The findings from Question 7 and Question 2 show reason to believe the price decreases. Insights from Question 1 and Question 4 give reason to believe that this is not always true without limitations.

H4: Complex models give significantly more accuracy in their predictions.

Based on findings from the literature research, the authors assumed that the more complex models would provide significantly more accuracy in their predictions, hence making the use of the simpler alternatives redundant. However, the authors were surprised to see that this was not the case in the present research.

Especially in the classification models predicting different scenarios, the more complex models RF and XGBoost outperform the simpler models. However, the differences are often solely marginal, not significant enough to justify the more complex model fitting procedure and longer run times especially for the needs of the mixed-role reseller.

Surprisingly, often the simple linear regression is amongst the best performing methods, also adding the advantage that the coefficient analyses provide insight into the individual impact of each variable. This is especially useful for the research on impact of characteristics of the sneakers on the resale price. Furthermore, the simplicity, interpretability and easy usage of the simpler methods cater especially to the needs of the mixed role reseller and were hence not redundant. The authors were further pleased to see that for Question 6, the simple regression tree provided good enough results and hence a plot of the actual tree could be added.

Concluding, the authors reject this hypothesis, as it cannot be said that for all models, the more complex models gave more accuracy and hence better predictions. Additionally, the differences in predictive accuracy were often minor.

H5: Celebrity and other collaborations increase the resale price.

Hypothesis 5, stating that celebrity collaborations, or collaborations in general, increase the resale price, was based on the author's assumption that any collaboration would make a sneaker more desirable. As can be seen in Chapter 1.1, in the history of sneaker development collaborations have always played an important role. This was also described in demand research papers (see Chapter 2.2.2).

In almost all scenarios and reseller questions highlighted previously, the binary variable *collab_b* has been of utmost significance. In the results of Question 1, Question 2, Question 4, Question 5, and Question 7, it was amongst the topmost important variables, indicating high significance in making investment decisions in the sneaker resale market.

Looking at the important variables of XGBoost in Question 1, for example, it can be seen how important the collaboration is in determining price premiums. Additionally, it can be seen when investigating the coefficients for the linear model that collaboration has a positive coefficient: this indicates that a collaboration will positively impact the buy decision of the reseller in this question, and therefore indicates that the collaboration leads to higher price premiums.

This also holds for a more long-term view, as can be seen in Question 5. The variable *collab_b* again has a positive coefficient and is significant for this model, indicating that a sneaker released with collaboration is more likely to provide increases in price premium, and hence price.

Taking a look at the coefficients for Question 7 as final example, one can see that the coefficient for collaboration is negative. In this scenario, this indicates that a sneaker released with collaboration will lead to a negative buy decision, which can be related to an increased price since the scenario aims for a low price premium. Of course, it cannot be finally concluded in this scenario that the negative “buy” decision is solely based on the collaboration. However, the negative coefficient as well as the high significance do suggest an important impact of this.

Concluding on all three examples, the authors find evidence in their research that collaboration positively impacts the price increase in the resale market. It is likely that a shoe released in collaboration will increase in price. However, due to the limitations e.g. of variable importance interpretation of XGBoost, the authors cannot conclude the extent of this impact. This means that although a significance can be identified, at this point the exact extent is not quantifiable. The price increases might be dependent on the collaboration, which cannot be analyzed with the data at hand. However, based on the underlying research, the authors accept Hypothesis 5.

7. Conclusion

The previous analysis focused on providing insights into the application of machine learning techniques to support data driven investment decisions in the sneaker resale market. After performing various analyses, focusing on disparate questions and scenarios the authors conclude that using these machine learning algorithms can be a very useful support in making investment decisions in the sneaker resale market.

The authors used the mentioned scenarios and questions, which two distinct reseller personas would presumably be highly interested in, in order to highlight how machine learning tools can support the very different decision-making procedures.

By putting emphasis on the different needs of the personas, keeping in mind their aims and overall goals of either large profits or collecting sneakers at lowest potential costs, the authors were able to highlight the advantages and disadvantages of each model.

The very practical approach chosen in this thesis, focusing to a great extent on applicability, allows sneakerheads, who engage in and are hence interested in information on resale markets, to profit from the insights.

The following summarizes the insights gained in relation to the research questions posed in Chapter 1.3.

RQ1: By what measures is the resale price of a sneaker impacted?

The various applications showed that both the prediction of the resale price itself, but also indirectly though the price premium, are impacted by the important variables of each model in each scenario. Observing and analyzing these gave meaningful insights. Taking Question 2 as an example, the overview of important variables showed that when using an XGBoost model to predict the resale price within the first 14 days after release, variables such as retail price, colors and different silhouettes increase the predictive power, and hence the accuracy of the predictions. This shows that using data including these characteristics of the shoes increases the predictions.

This can be further observed in Question 3, where the coefficients of the linear model predicting the maximum price premium to be achieved within the first 14 days after release gave insight into how these characteristics changed the price premium prediction, which relates to the price prediction. Here, it was clear to see the coefficient for retail price was negative, leading to a lower predicted price premium for shoes that were more expensive from the beginning. Additionally, the model coefficients showed the very positive impact of one specific silhouette on the maximum price premium, indicating that buying a shoe of this kind would with high probability increase in maximum price premium over the observed time frame.

Question 7 also highlights how the resale price is impacted, by showing for example that a shoe released in collaboration will decrease the probability for a reseller to buy it, if this reseller aims to pay the lowest resale price possible. This gives a clear indication that a shoe released in collaboration will increase in resale value.

Overall, each of these examples as well as the thorough analysis and discussion chapters before, show that the resale price of a sneaker is impacted by its characteristics. These are relevant findings for RQ1, exploring by what measures the resale price of a sneaker is impacted. Coefficient estimates and important variable measures of the models are a good tool to examine in which way this occurs. However, it is important to realize that the impact of each characteristic depends on the scenario and question that is being investigated.

RQ2: ***How can machine learning techniques help sneakerheads in their decision to buy a sneaker with the intention to resale at profit?***

> **RQ2a:** *Which machine learning techniques provide the most accurate predictions for professional resellers investment decisions?*
>
> **RQ2b:** *Which machine learning techniques provide simple, interpretable models, and yet result in a decent level of accuracy for mixed-role investment decisions?*

The authors can conclude that the application of machine learning methods can be useful to make more founded, data-driven investment decisions. The research on the basis of applying these measures to different scenarios and questions showed how the different methods can be useful in varying settings and situations. Making use of the classification algorithms as well as regression tasks allowed for more variance in the decision making. No one method continuously outperformed the others - the combination of several methods and selection of those that best performed according to the respective needs of the personas seemed to be the best approach.

It could be seen that for the professional reseller, although accuracy was the main criteria, the most complex and hence expectedly most accurate models were not always the best choice. In the two classification questions, Question 1 and Question 4, the highest accuracies were on average achieved with either RF or XGBoost. However, the simple linear regression model did not perform significantly worse, which gave enough justification to analyzing the related coefficient estimates for further insight. For the two regression problems, Questions 2 and Question 3, the best models were XGBoost and linear regression.

Hence, for the scenarios made for the professional reseller, one can conclude that the most accurate predictions were on average gained with complex models such as RF and XGBoost, however the simpler models did not provide significantly worse results. For the professional reseller, who is interested in accuracy but also in determining the influencing factors of the price development, a combination of both more complex and simpler models would be most advisable.

For the mixed-role reseller, the research showed that the simple models linear and logistic regression and KNN deliver good results and provide satisfactory accuracy. Seeing that the main aim of the mixed-role reseller is not high accuracy, but simple, quick and easily understandable results, mixed-role resellers can rely on these simpler models for their purchase decisions.

RQ3: How do celebrity and other collaborations impact the resale price?

The final research question aimed to find out how celebrity and other collaborations impact the resale price. This is especially interesting for the sneaker resale market, as, based on the insights gained from the literature research, sneakerheads put a lot of focus on group belonging, following their heroes, and wearing shoes that have some sort of reputation.

After having performed the analysis, and evaluated both important variables and, if applicable, coefficient estimates, the authors can conclude that collaborations do impact resale price significantly. The collaboration variable appeared frequently in the most important variables, indicating its importance for highly accurate models. Additional insights from the coefficient estimates, where applicable, gave insight into the positive influence a collaboration has on the resale value.

In Question 1, as an example, the outcome *buy* is predicted in the case of a positive price premium, meaning an increase in price. The coefficient estimate for collaboration in this case is positive, indicating a correlation between a collaboration and an increased price.

This also holds vice versa, as seen in Question 7. Here, the decision predicted is to *buy* at a preferably low price premium. In this case, the coefficient estimate for *collab_b* is negative, indicating that should the independent variable increase, the dependent variable decreases. This shows that the larger the value *collab_b* is, the smaller the potential of the reseller buying the shoe due to the price being too high for the reseller.

The authors conclude therefore that the existence of a collaboration has a positive impact on the resale value of a sneaker. However, the extent of this positive influence, and if different collaboration partners would influence the price differently, cannot be determined based on this research.

Limitations and further research

The authors are well aware that their research cannot be seen as all-encompassing and is limited by several aspects. These will be highlighted in the following and can provide interesting opportunities for further research.

One limiting factor is the data used. Since the data was taken from one resale platform only, the modelling, analysis and conclusions drawn hence can only be related to the platform StockX. Hence, the findings cannot be applied unconditionally across all other resale platforms and markets, since also unofficial resale markets, where a buyer and seller negotiate directly, exist. The findings from this paper can be used as indication for e.g. price influencers, but not all-encompassing to all resale sites.

Additionally, the data included only deadstock data over a limited time frame of six months and only data for the three brands adidas, Nike and Air Jordan, which also limits the research. This means the findings cannot be applied without limitations to sneakers that are not included within these data specifications.

One of the most prominent limitations, which also provides a myriad of future research options, is the incorporation of non-quantitative price influencers such as hype in the modelling. Based on the findings from the literature review, external factors such as hype, communication about a shoe and release events are of major importance in raising awareness for sneakers. It can be assumed that resale price development is influenced by this. However, in the data and modelling in this thesis, hype could not be included. In Question 2, it could be seen that the data used does not suffice entirely for resale price predictions. The models had large prediction errors and not optimal performance. Hence, one can hypothesize that additional factors significantly influence these decisions. A starting point for further research could be to perform text analysis on tweets posted referring e.g. to new releases and analyze the sentiment as well as numbers of retweets and likes. The authors aimed to add this research, however decided to refrain from doing so. The data gathering of tweets from twitter has posed a problem, since the free accounts do not provide unlimited data usage and crawling.

Additionally, the extent to which collaborations impact the price could not be analyzed in this setting. The authors can conclude that a collaboration will have impact on the resale price - the extent of this impact however, as well as how this depends on the respective person, company or group used in the collaboration, could not be identified with the present data. A further research option could be to analyze how different personas (e.g. sports stars vs. fashion celebrities) impact the resale value of a sneaker or if there is an advantage in choosing a collaboration with a single person vs. a collaboration with another brand.

Final Remarks

Concluding the thesis, the authors summarize that machine learning techniques can be helpful in supporting investment decisions in the sneaker resale market. Using both simple and more complex models and applying these to clear scenarios and key questions enabled the authors to identify price driving characteristics of sneakers and showed how using these techniques can allow sneakerheads to “level up their sneaker game”.

References

Adebiyi, A. A., Adewumi, A. O., & Ayo, C. K. (2014). Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction. *Journal of Applied Mathematics*, *2014*, 1–7. https://doi.org/10.1155/2014/614342

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, *147*, 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038

Argo, J. J., Dahl, D. W., & Morales, A. C. (2006). Consumer Contamination: How Consumers React to Products Touched by Others. *Journal of Marketing*, *70*(2), 81–94. http://www.jstor.org/stable/30162087

Britannica, T. Editors. (2019). Supply and demand. In *Encyclopedia Britannica*. https://www.britannica.com/topic/supply-and-demand

Brock, T. C. (1968). Implications of Commodity Theory for Value Change. *Psychological Foundations of Attitudes*, 243–275. https://doi.org/10.1016/b978-1-4832-3071-9.50016-7

Cassidy, N. G. (2018). *The Effect of Scarcity Types on Consumer Preference in the High-End Sneaker Market* [Honors Thesis]. https://libres.uncg.edu/ir/listing.aspx?id=23023

Chai, J., Wang, Y., Wang, S., & Wang, Y. (2019). A decomposition–integration model with dynamic fuzzy reconstruction for crude oil price prediction and the implications for sustainable development. *Journal of Cleaner Production*, *229*, 775–786. https://doi.org/10.1016/j.jclepro.2019.04.393

Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. *Journal of Computational and Applied Mathematics*, *365*, 112395. https://doi.org/10.1016/j.cam.2019.112395

Choi, J. W., Cluver, B., & Kim, M. (2015). Who Are These Sneakerheads? *International Textile and Apparel Association Annual Conference Proceedings*, *72*(1). https://www.iastatedigitalpress.com/itaa/article/id/2731/

Choi, Ji. W. (2017). *Sneakerheads' Assessment of Sneaker Value and Behaviors throughout the Sneaker Ownership Cycle*. Ir.library.oregonstate.edu. https://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/9306t3488

Chu, H. (2013). A conceptual model of motivations for consumer resale on C2C websites. *The Service Industries Journal*, *33*(15-16), 1527–1543. https://doi.org/10.1080/02642069.2011.636422

Chu, H., & Liao, S. (2007). Exploring Consumer Resale Behavior in C2C Online Auctions: Taxonomy and Influences on Consumer Decisions. *Academy of Marketing Science Review*, *11*(3).

Chu, H., & Liao, S. (2010). Buying while expecting to sell: The economic psychology of online resale. *Journal of Business Research*, *63*(9-10), 1073–1078. https://doi.org/10.1016/j.jbusres.2009.03.023

Cochrane, L. (2014). *Supreme's shoe collaboration with Nike provokes a "riot."* The Guardian. https://www.theguardian.com/fashion/fashion-blog/2014/apr/04/supreme-nike-shoe-collaboration-provokes-riot-foamposite

Dayton, S. (2020). *HOW TO RESELL SNEAKERS: BEGINNER'S GUIDE TO FLIPPING SHOES AND MAKING REAL MONEY*. Nichepursuits.com. https://www.nichepursuits.com/reselling-shoes/

Denny, I. (2020). The sneaker – marketplace icon. *Consumption Markets & Culture*, 1–12. https://doi.org/10.1080/10253866.2020.1741357

Díaz, G., Coto, J., & Gómez-Aleixandre, J. (2019). Prediction and explanation of the formation of the Spanish day-ahead electricity price through machine learning regression. *Applied Energy*, *239*, 610–625. https://doi.org/10.1016/j.apenergy.2019.01.213

Dolbec, P.-Y., & Parmentier, M.-A. (2019). Believe the Hype: How Resale Monetized Desire. *NA - Advances in Consumer Research*, *47*, 539–540. Association for Consumer Research. https://www.acrwebsite.org/volumes/2551344/volumes/v47/NA-47

Dvorkin, Y., Wang, Y., Pandzic, H., & Kirschen, D. (2014). Comparison of scenario reduction techniques for the stochastic unit commitment. *2014 IEEE PES General Meeting | Conference & Exposition*. https://doi.org/10.1109/pesgm.2014.6939042

Fromkin, H. L. (1970). Effects of experimentally aroused feelings of undistinctiveness upon valuation of scarce and novel experiences. *Journal of Personality and Social Psychology*, *16*(3), 521–529. https://doi.org/10.1037/h0030059

Gaillard, P., Goude, Y., & Nedellec, R. (2016). Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. *International Journal of Forecasting*, *32*(3), 1038–1050. https://doi.org/10.1016/j.ijforecast.2015.12.001

Gierl, H., & Huettl, V. (2010). Are scarce products always more attractive? The interaction of different types of scarcity signals with products' suitability for conspicuous consumption. *International Journal of Research in Marketing*, *27*(3), 225–235. https://doi.org/10.1016/j.ijresmar.2010.02.002

Hastie, T., Tibshirani, R., Friedman, J., & Springerlink (Online Service. (2009). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer New York.

Herrera, G. P., Constantino, M., Tabak, B. M., Pistori, H., Su, J.-J., & Naranpanawa, A. (2019). Long-term forecast of energy commodities price using machine learning. *Energy*, *179*, 214–221. https://doi.org/10.1016/j.energy.2019.04.077

Hovland, C. I., & Weiss, W. (1951). The Influence of Source Credibility on Communication Effectiveness. *Public Opinion Quarterly*, *15*(4), 635. https://doi.org/10.1086/266350

Hwang, Y., Ko, E., & Megehee, C. M. (2014). When higher prices increase sales: How chronic and manipulated desires for conspicuousness and rarity moderate price's impact on choice of luxury brands. *Journal of Business Research*, *67*(9), 1912–1920. https://doi.org/10.1016/j.jbusres.2013.11.021

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

Kohzadi, N., Boyd, M. S., Kermanshahi, B., & Kaastra, I. (1996). A comparison of artificial neural network and time series models for forecasting commodity prices. *Neurocomputing*, *10*(2), 169–181. https://doi.org/10.1016/0925-2312(95)00020-8

Kramer, L. (2019). *How Does the Law of Supply and Demand Affect Prices?* Investopedia. https://www.investopedia.com/ask/answers/033115/how-does-law-supply-and-demand-affect-prices.asp#:~:text=It

Kuhn, M., Johnson, K., & Springer Science+Business Media. (2016). *Applied predictive modeling*. Springer.

Lessmann, S., & Voß, S. (2017). Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy. *International*

Journal of Forecasting, *33*(4), 864–877. https://doi.org/10.1016/j.ijforecast.2017.04.003

Lynn, M. (1989). Scarcity effects on desirability: Mediated by assumed expensiveness? *Journal of Economic Psychology*, *10*(2), 257–274. https://doi.org/10.1016/0167-4870(89)90023-8

Ma, K., & Treiber, M. C. (2020). *Hedonic Pricing in the Sneaker Resale Market* [Msc Thesis]. https://sites.duke.edu/djepapers/files/2020/08/matreiber-dje.pdf

Maher, T. (2019). *"The Sneaker Game" A Documentary*. Www.youtube.com. https://www.youtube.com/watch?v=iXae9pMBcR4

Marshall, A. (1920). *Principles of economics: an introductory volume* (8th ed.). Macmillan And Co.

McCracken, G. (1986). Culture and Consumption: A Theoretical Account of the Structure and Movement of the Cultural Meaning of Consumer Goods. *Journal of Consumer Research*, *13*(1), 71–84. https://doi.org/10.1086/209048

Mulligan, R. F., & Grube, A. J. (2006). Modelling Markets for Sports Memorabilia. *Journal of Economics and Economic Education Research*, *7*(2), 75–102. https://ssrn.com/abstract=1028843

Neap, H. S., & Celik, T. (1999). Value of a Product: A Definition. *International Journal of Value-Based Management*, *12*(2), 181–191. https://doi.org/10.1023/a:1007718715162

Norman, L. (2020). *Predicting StockX Sneaker Prices With Machine Learning*. Medium. https://medium.com/swlh/predicting-stockx-sneaker-prices-with-machine-learning-ec9cb625bec0

Nowotarski, J., & Weron, R. (2014). Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, *30*(3), 791–803. https://doi.org/10.1007/s00180-014-0523-0

Pal, N., Arora, P., Kohli, P., Sundararaman, D., & Palakurthy, S. S. (2018). How Much Is My Car Worth? A Methodology for Predicting Used Cars' Prices Using Random Forest. *Advances in Intelligent Systems and Computing*, 413–422. https://doi.org/10.1007/978-3-030-03402-3_28

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*, *42*(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040

Pudaruth, S. (2014). Predicting the Price of Used Cars using Machine Learning Techniques Predicting the Price of Used Cars using Machine Learning Techniques. *International Journal of Information & Computation Technology*, *4*(7), 753–764.

Rico-Juan, J. R., & Taltavull de La Paz, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems with Applications*, *171*, 114590. https://doi.org/10.1016/j.eswa.2021.114590

Seiler, R., & Kucza, G. (2017). Source Credibility Model, Source Attractiveness Model And Match-Up-Hypothesis–An Integrated Model. *Economy & Business, Journal of International Scientific Publications*, *11*(1), 1–15.

Semmelhack, E., Garcia, B., Lepri, E., Willis, P., & Hatfield, T. (2015). *Out of the Box: The Rise of Sneaker Culture*. Rizzoli Electa.

Servantes, I. (2021). *This is why you can't get the hyped Nike and Jordan sneakers you want*. Input. https://www.inputmag.com/style/nike-air-jordan-1-trophy-room-snkrs-resell-backdoor-bots

Shah, S. (2019). *Predicting Sneaker Resale Prices using Machine Learning* [Thesis]. https://repository.asu.edu/items/52272

Steinberg, L. (2018). *The Profitable Hidden Sneaker Market*. Forbes. https://www.forbes.com/sites/leighsteinberg/2018/09/17/the-profitable-hidden-sneaker-market/?sh=fce3d5a59256

Sweeney, J. C., & Soutar, G. N. (2001). Consumer perceived value: The development of a multiple item scale. *Journal of Retailing*, *77*(2), 203–220. https://doi.org/10.1016/s0022-4359(01)00041-0

ThredUp. (2020). *2020 Fashion Resale Market and Trend Report*. Thredup.com. https://www.thredup.com/resale/

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. *Procedia Computer Science*, *174*, 433–442. https://doi.org/10.1016/j.procs.2020.06.111

Tsjeng, Z. (2014). *The Supreme x Nike launch that almost caused a riot*. Dazed. https://www.dazeddigital.com/fashion/article/19459/1/nypd-shut-down-supreme-nike-sneakers-launch-store-public-safety-riot-concerns

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tynan, C., McKechnie, S., & Chhuon, C. (2010). Co-creating value for luxury brands. *Journal of Business Research*, *63*(11), 1156–1163. https://doi.org/10.1016/j.jbusres.2009.10.012

Wade, R. (2021). *The global sneaker resale market could reach $30 billion by 2030*. Finance.yahoo.com. https://finance.yahoo.com/news/global-sneaker-resale-market-could-reach-30-billion-by-2030-cowen-191003371.html xgboost developers. (n.d.). *Understand your dataset with XGBoost — xgboost 1.5.0-dev documentation*. Xgboost.readthedocs.io. Retrieved May 27, 2021, from https://xgboost.readthedocs.io/en/latest/R-package/discoverYourData.html

Zhang, T. (2020, September 22). *Predicting Sneaker Resell With Deep Learning*. Medium. https://medium.com/swlh/predicting-sneaker-resell-with-deep-learning-d3a78b144099