Decoding Consumer Preferences in Wine: Predictive Analytics and Machine Learning in Analyzing Portuguese Wine Consumer Ratings

Lukas Florian Schneider

Dissertation written under the supervision of Professor Nicolò Bertani

Dissertation submitted in partial fulfillment of requirements for the MSc in Business Analytics, at the Universidade Católica Portuguesa, 3rd January 2024.

Abstract

Title: Decoding Consumer Preferences in Wine: Predictive Analytics and Machine Learning in Analyzing Portuguese Wine Consumer Ratings

Author: Lukas Florian Schneider

The increasing significance of electronic markets and platforms has revolutionized how consumers interact with and purchase products such as wine. This change provides more information availability to distinguish it from others. Consequently, it is easier to convey a distinct message for characterful products. Nevertheless, it also results in drawbacks like information overload and a loss of the ability to differentiate. This thesis uses advanced analytics to decode the most important factors for consumer preferences in the Portuguese wine market. At the same time, it addresses the challenges and opportunities presented by the information-rich environment of electronic marketplaces.

Specifically, I conducted a study to identify, using predictive analytics tools, the essential qualitative product features of Portuguese wine that matter for consumer satisfaction.

To do this, robust predictive models were built using individual consumer review ratings and the descriptive characteristics of Portuguese wines from the electronic marketplace and platform Vivino. Ultimately, relative feature importance was determined using Random Forest, AdaBoost, Gradient Boosting, and XGBoost. Additionally, the study incorporates user comments via topic modeling into the predictive models.

As a result of this analysis, the study revealed consumer participation, user engagement, sensory perception, generalization, and price description as driving factors for consumer satisfaction.

In summary, all models demonstrated similar outcomes, recommending a focus on extrinsic rather than intrinsic product attributes to differentiate from other product groups. These findings can be used further for strategic market decisions and research.

Keywords: Consumer Satisfaction, Wine Analytics, Predictive Modeling, Topic Modeling, Feature Importance

Abstract Portuguese

Title: Title: Decoding Consumer Preferences in Wine: Predictive Analytics and Machine Learning in Analyzing Portuguese Wine Consumer Ratings

Author: Lukas Florian Schneider

A crescente importância dos mercados e plataformas eletrônicos revolucionou a forma como os consumidores interagem e compram produtos, especialmente o vinho. Esta mudança oferece maior disponibilidade de informação para destacar produtos únicos, mas também traz desafios como a sobrecarga de informação e perda de diferenciação. Esta tese investiga o uso de análises avançadas para decifrar os principais fatores que influenciam as preferências dos consumidores no mercado de vinhos português, abordando os desafios e oportunidades do ambiente rico em informações desses mercados.

Realizei um estudo que emprega ferramentas de análise preditiva para identificar as características qualitativas essenciais do vinho português que impactam a satisfação do consumidor. Modelos preditivos robustos foram desenvolvidos usando avaliações de consumidores e características dos vinhos portugueses em mercados eletrônicos e na plataforma Vivino. Métodos como Random Forest, AdaBoost, Gradient Boosting e XGBoost ajudaram a determinar a importância relativa dessas características, incorporando também comentários dos usuários através da modelagem de tópicos.

Os resultados revelaram que a participação e envolvimento do consumidor, percepção sensorial, generalização e descrição de preço são fatores cruciais para a satisfação do cliente. Todos os modelos apontaram para a necessidade de focar em atributos extrínsecos, ao invés de intrínsecos, para se diferenciar no mercado. Essas descobertas são valiosas para estratégias de mercado e pesquisas futuras.

Keywords: Consumer Satisfaction, Wine Analytics, Predictive Modeling, Topic Modeling, Feature Importance

Contents

1 Introduction 1

2 Background 3

2.1 Context 3

2.1.1 World of Wine in Portugal 3

2.1.2 Vivino as Global Marketplace 5

2.2 Review of relevant literature 5

2.2.1 Customer Satisfaction 5

2.2.2 Leveraging Machine Learning's Feature Importance 7

2.3 Classification with Ensemble Models 8

2.3.1 Random Forest 10

2.3.2 AdaBoost 11

2.3.3 Gradient Boosting 12

2.3.4 XGBoost 13

3 Data 14

3.1 Data Extraction 15

3.2 Data Preprocessing 16

3.3 Topic Modeling 18

3.4 Descriptive and Exploratory Data Analysis 22

4 Results 26

4.1 Modeling 27

4.2 Model Evaluation & Business Impact 29

4.2.1 Random Forest 29

4.2.2 AdaBoost 31

4.2.3 Gradient Boosting 34

4.2.4 XGBoost 37

4.3 Business Insights 41

5 Conclusion 44

6 Bibliography 47

List of Figure

Figure 1: Comparative Illustration of Bagging vs. Boosting in Ensemble Learning................. 9
Figure 2: Illustration of Random Forest Ensemble Technique 11
Figure 3: Illustration of AdaBoost Ensemble Technique... 12
Figure 4: Illustration of Gradient Boosting Ensemble Technique 13
Figure 5: Vivino Consumer 'review_rating' Example... 15
Figure 6: Vivino 'ratings_average' for Wine Vintage Example 15
Figure 7: Graphical Comparison of Original LDA vs. LDA Approximation Method 20
Figure 8: Coherence Score Across Different Topic Numbers for LDA Model 21
Figure 9: Number of consumer review ratings (2012-2023)... 23
Figure 10: Distribution of Region, Wine types, and Comment Topics in the Wine Dataset... 24
Figure 11: Distribution Comparison for 'review_rating' before and after classification 26
Figure 12: Random Forest Feature Importances by Product Attribute Category 30
Figure 13: AdaBoost Feature Importances by Product Attribute Category 33
Figure 14: Gradient Boosting Feature Importances by Product Attribute Category................ 36
Figure 15: XGBoost Feature Importances by Product Attribute Category 39
Figure 16: Illustration of Reviewer Rank... 42
Figure 17: Possible scale for comparing the price with the average rating.............................. 42
Figure 18: Possible improvements for a more effective wine description............................... 44

List of Tables

Table 1: Overview of Vineyard Areas and Yearly Wine Production by Region in Portugal (2022) 4
Table 2: Product Attribute Categorization for Gathered Wine Data....................................... 18
Table 3: Top Words Representing the Five Main Topics Derived from Wine Reviews 22
Table 4: Distribution of Numerical Values in the Wine Dataset 25
Table 5: Comparative Performance Metrics of Ensemble Learning Models........................... 28

List of Equations

Equation 1: Calculation of Feature Importance for Feature 8
Equation 2: XGBoost Regularized Loss Function Formulation 14

List of Abbreviations

API	Application Programming Interface
Br-PT	Brazilian Portuguese
DOC	Denominação de Origem Controlada
IPR	Indicação de Proveniência Regulamentada
KPI	Key Performance Indicator
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
N.V	Non-Vintage
Pt-PT	Portugal Portuguese
ROC-AUC	Receiver Operating Characteristic – Area Under Curve
SEO	Search Engine Optimization
SVD	Singular Value Decomposition

Acknowledgment

First and foremost, I express my gratitude to my family for their support throughout my academic journey. Their encouragement and motivation have been invaluable. I am thankful for their presence in my life, which has enabled me to pursue my academic goals with confidence and determination.

I owe an immense debt of gratefulness to my advisor, professor Nicolò Bertani, for his exceptional guidance and support. His advice on both my thesis and career choices has been invaluable to my academic success. I appreciate his willingness to provide guidance and the meaningful insights that have shaped my professional path. I would like to thank you for your mentorship and belief in my abilities during the whole time at Católica Lisbon School of Business.

Sincere thanks to my friends and roommates, whose unwavering belief in me and constant motivation have been a source of immense support. Their encouragement and faith in my abilities have been invaluable during this journey.

I acknowledge the use of DeepL and OpenAI's ChatGPT for specific language improvements in this thesis, including refining word choice, sentence structure, and correcting grammar based on provided prompts. The intellectual content and original ideas, however, are solely my own. Should there be interest, I am willing to provide the specific prompts used in this process.

Finally, I have to thank the developer of the basic structure of Vivino API for publishing it on Stack Overflow. This work could not have been done without them.

1 Introduction

As the internet revolutionizes access to information, it becomes crucial to understand its impact on consumer perception. This is especially important in the context of the unique and historically rich Portuguese wine industry. Greater accessibility allows for clearer communication of key product attributes (Choudhury et al., 1998) and can help differentiate wines from other countries (Menghini, 2015). Electronic marketplaces and digital platforms boost this evolution (Wang et al., 2008). In addition, they can potentially enhance the distinct product experience by adding another layer with the inclusion of personal comments and reviews (Simonson, 2015). However, this advantage is not without its drawbacks as the overwhelming amount of information can confuse consumers (Grover et al., 2006) and hinder the effective communication of the intended message. Therefore, it is crucial to comprehend the most pertinent information.

For this purpose, this study aims to identify the most important characteristics of Portuguese wine information. To accomplish this objective, machine learning algorithms and statistical methods were utilized to predict consumer satisfaction. The satisfaction was evaluated based on individual consumer review ratings on a five-star scale (Westbrook, 1980). This investigation deciphers consumer preferences in the information-rich landscape of electronic markets while dealing with the vast amount of information presented in such areas. Furthermore, this research indicates that effective information management involves not only simplification but also understanding and predicting consumer preferences.

To create robust predictive models, this study integrates consumer reviews and wine attributes. A literature-based framework categorizes these attributes, distinguishing between features that are directly and indirectly related to Portuguese wine, as well as personal experiences reflected in the comments. From the Danish wine database and the electronic marketplace Vivino, a rich dataset specific to Portuguese wine was gathered, comprising both qualitative and quantitative data. This dataset, including 80,021 ratings and 73 product attributes (Vivino, 2023b) was analyzed to determine the feature most influential in review ratings. This approach helps to avoid overwhelming information and ensures the conveyance of key messages. The insights gained are valuable for various stakeholders. They can improve the presentation of information on a webpage, as seen on Vivino. Furthermore, they can assist in developing future production plans for wine producers. Ultimately, they can also aid in creating market strategies for specific regions or countries. In today's competitive wine market, characterized by more demanding consumers and an increasing number of wine producers (Faria et al., 2020), focusing on the most relevant information is crucial. This is particularly true for traditional products such as Portuguese wine, to maintain the ability to distinguish.

To analyze these essential aspects of Portuguese wine, the comments associated with the consumer review rating were incorporated by applying topic modeling using Latent Dirichlet Allocation. For the predictive analytics component, a classification into nine different categories of the specific consumer review ratings was performed using Random Forest, AdaBoost, Gradient Boosting, and XGBoost. Each model's unique strengths were leveraged to handle the complex patterns of customer satisfaction and to build robust models (refer to chapters 2.3 and 4.2 for further information). To determine the desired outcome, namely the relevance of the product attributes for Portuguese wine, the importance of the attributes was calculated for each model. The importance scores were divided into product categories based on the literature to obtain more detailed consumer insights. This allowed the most important product category to be assessed. Although the four models used different approaches to classify consumer ratings, a common overall picture and trend for information areas could be observed. The study identified several key components, including consumer participation, user engagement, sensory perception, price description, and generalization. Based on these findings, actionable recommendations were developed to enhance the presence of Portuguese wine on wine platforms such as Vivino. This study culminates in strategic recommendations tailored to enhance the digital footprint of Portuguese wine, to convey the distinct message of Portuguese wine, and to foster consumer engagement.

The following chapters provide a structured overview of this study. The first chapter is background about Portuguese wine to emphasize the importance of why a message needs to be conveyed and about Vivino and why it can help to do so. After reviewing the literature on the product attributes that drive customer satisfaction, the four machine-learning tools were introduced and applied to the web-scraped data from Vivino by leveraging their distinct characteristics. The resulting analysis was then used to derive business impacts and recommendations to maximize the potential of the provided information by focusing on the most significant attributes. In the concluding chapter, the study's limitations are presented to provide implications for future research or the application of this study in the future.

2 Background

To understand how predictive analytics can assess the most important features of information for Portuguese wine on Vivino, this chapter establishes the background for the underlying key components: Portuguese wine, customer satisfaction, and machine learning models. The initial part briefly introduces Portuguese wine and its unique features. Subsequently, Vivino, an electronic marketplace and platform for wine, is introduced as the source of data. After examining specific product attribute categories for consumer satisfaction based on a literature review, a link between this satisfaction and machine learning models is established. Ultimately, this chapter also introduces the models used in this study, providing a brief evaluation of their effectiveness in accurately predicting consumer review ratings.

2.1 Context

2.1.1 World of Wine in Portugal

Machine Learning and Statistical Concepts represent not only the cutting edge of technology (Chen & Guestrin, 2016) but also a bridge from tradition to modern analysis. By leveraging a spectrum of product attributes, strongly affected by history, such methods enable informed predictions. This study connects these advanced analytics tools with one of the world's oldest beverages, wine (Mayson, 1998), and one of Europe's oldest wine cultivators, Portugal (Robinson, 2006). In Portugal, wine is not just a beverage but a key part of the country's historical fabric. Understanding its history is essential to grasp the enduring characteristics of Portuguese wine, as many historical elements remain influential today and drive a variety of product attributes.

Firstly, wine cultivation was established by the Tartessians in the Tagus area around 2000 BC. Over time, grape variety and production techniques were expanded and extended by the Phoenicians (Phillips, 2002). This evolution continued with the Ancient Greeks, Celts, and Romans, who spread wine cultivation throughout the country (Mayson, 1998). The subsequent centuries were shaped by the influence of various cultures that held positions in the Iberian Peninsula, notably the Arabs and Christians during the Reconquista (Phillips, 2002). Furthermore, the voyages conducted by Portuguese explorers further contributed to the diversification of wines (Mayson, 1998). By the 17th century, Portugal had become a significant wine exporter, particularly noted for its fortified wines like Port and Madeira wine, which largely traded with England (Bradford, 1983). Despite devastation in the late 19th century due to a French invasion, the Portuguese wine market recovered, aided once again by

England. Due to this prestigious history, Portugal is considered a part of the old world of wine (Mayson, 1998). These wines were tried and tested over centuries and maintained, whereas the wine of the new world (e.g., from California) was produced mainly by modern consumers (Robinson, 2006). Consequently, Portuguese winemakers compete with these (Menghini, 2015) and are faced with the modern aspects of consumer demands.

In recent years, Portugal's integration into the European Union has notably influenced wine cultivation. This accession mainly involved aligning local wine laws with EU standards. As a result quality designations like *Denominação de Origem Controlada (DOC)* and *Indicação de Proveniência Regulamentada (IPR)* were established (Mayson, 1998). Despite these regulatory changes, the unique characteristics of Portuguese wine remain evident. The country's wine world today still mirrors these specifications and historical events, especially in the impressive variety and spread of grape and taste varieties. Portugal boasts nearly 500 different types, distributed across the mainland and islands like the Azores and Madeira (Robinson, 2006). This wide array of grapes is a testament to Portugal's rich viticultural history, a fact that is further illustrated by vineyards covering two percent of the nation's land area. Table 1 indicates that this occupation is widespread throughout the country, encompassing both vineyards and production for the year 2022 (Instituto da Vinha e do Vinho, 2023). Additionally, it strengthens the importance of wine production for the whole country's economy.

Region	**Vineyards (hectare)**	**Yearly Production (hectoliter)**
Alentejo	25,924	1,134
Algarve	1,442	16
Beira Atlântico	13,391	202
Douro e Porto	43,466	1,457
Lisboa	19,869	1,194
Minho	24,371	1,004
Península de Setúbal	8,027	500
Tejo	12,847	686
Terras da Beira	14,059	202
Terras de Cister	2,274	48
Terras do Dão	13,409	279
Trás-os-Montes	9,701	83
Açores	1,709	5
Madeira	681	38
Total	**191,170**	**6,848**

Table 1: Overview of Vineyard Areas and Yearly Wine Production by Region in Portugal (2022), data from: (Instituto da Vinha e do Vinho, 2023)

Contrary to the production forecast for most wine producers worldwide, Portugal is expected to experience growth in 2023 (International Organisation of Vine and Wine, 2023a). This increase will also strengthen its current position as the 10th largest wine producer and 8th largest wine exporter worldwide (International Organisation of Vine and Wine, 2023b).

2.1.2 Vivino as Global Marketplace

This rich historical context sets the stage for understanding how Portuguese wine, with its deep-rooted legacy, navigates the modern challenges of a globalized market. In this landscape, electronic marketplaces like Vivino play a pivotal role. These marketplaces provide a critical resource to exhibit Portugal's exceptional products and efficiently compete in today's interconnected global wine world (Wang et al., 2008).

Vivino, a Danish company, operates a website and mobile app that offers a vast of information on wine attributes. In addition, this platform enables users to rate, comment on, search for, and review wines. It is currently recognized as the largest wine platform worldwide and the most downloaded wine app, boasting a constantly expanding database. Currently, it includes 17 million listed wines across 3,477 wine regions and has approximately 66 million users. The company reports 283 million ratings and 98 million reviews (Vivino, 2023a). In addition to these features, Vivino provides services like identifying the cheapest online prices for wines and the option to purchase them directly from the platform. The app utilizes gathered data about wines and users to generate personalized wine recommendations (Marr, 2021).

Specifically for Portugal, Vivino lists 22,827 wines from 3,218 wineries (Vivino, 2023b). Thus, Vivino's extensive database, including its qualitative and comment features, can be leveraged to quantify customer satisfaction with Portuguese wine. Consequently, it aids in determining the most important information for consumers.

2.2 Review of relevant literature

2.2.1 Customer Satisfaction

Furthermore, electronic marketplaces like Vivino provide the possibility to assess the product's quality. This assessment is fundamental, as it directly influences customer satisfaction and the probability of making initial and subsequent purchases, thereby affecting profitability (Aaker & Jacobsen, 1994). This study focuses on customer satisfaction, employing a transaction-specific approach to evaluate consumer contentment post-purchases (Oliver, 1977). Relying on this definition, customer satisfaction was measured using consumer star ratings (Westbrook,

1980) for an individual consumer review rating, each accompanied by a comment. Additionally, the direct link between customer satisfaction and product characteristics (Andaleeb & Conway, 2006) was leveraged to determine the most important features. To facilitate deeper predictive analytics using machine learning models, product characteristics were categorized as extrinsic, intrinsic, and personal.

Among these, intrinsic product attributes, are closely intertwined with sensory perception. They include essential attributes like scent, flavor, noise, consistency, and mouthfeel (Enneking et al., 2007). Regarding wine, they are associated with the experience of drinking (Robertson et al., 2018). The significance of these attributes is a subject of debate in the literature. While some researchers argue that they significantly affect customer value, market participation, and satisfaction (Levitt, 1983), others maintain that their impact is less substantial (Carpenter et al., 1994).

In contrast to intrinsic attributes, extrinsic attributes do not directly link to the product itself (Symmank, 2018). As they are not inherent to the product, consumers cannot experience them firsthand, which leads in turn to indeterminacy (Akdeniz et al., 2013). Hence, customers search for them and rely on available information to make informed purchasing decisions (Nelson, 1970). This tendency underscores the pivotal role of electronic markets in providing such product information (Choudhury et al., 1998). In the context of wine, consumers can determine these attributes prior to purchasing a bottle of wine, independent of the wine's quality (Horowitz & Lockshin, 2002). For instance, price, age, which in the context of wine is referred to as vintage, brand, and region of origin are categorized as this type of product attribute (Robertson et al., 2018). The literature shows disagreement regarding the importance of these attributes. Some studies consider them as complementing intrinsic values (Brechan, 2006), whereas others contend that they are crucial (Richardson et al., 1994). The growth of the internet demands a reassessment of the functions of these characteristics (Rao & Monroe, 1989). Furthermore, when sensory perception is lacking, consumers frequently rely on extrinsic factors when making decisions (Akdeniz et al., 2013). Nonetheless, it is important to note that there may be counterfactual effects between these two kinds of attributes (Muthukrishnan & Kardes, 2001).

However, the aforementioned intrinsic and extrinsic attributes do not fully account for the influence of peer evaluations, which play a crucial role in shaping individual decision-making processes. The evaluation and perspectives of relevant peer groups can significantly impact consumer decisions (Kim et al., 2014). To capture this dimension in the predictive machine learning models in this study, user comments and associated characteristics, such as experience and cultural background, were used. This integration is vital for a comprehensive understanding of consumer preferences and overall satisfaction. By conducting a joint analysis that combines both objective attributes and hidden dimensions, such as personal perceptions and peer influences (Luo et al., 2008). Consequently, the models are better equipped to elucidate the nuances of consumer choice and are more robust. This approach acknowledges the complexity of consumer behavior, where personal perceptions, influenced by peer feedback and cultural context, are integral to understanding the diversity in consumer preferences.

Nevertheless, all attributes share a central characteristic: they are considered information. In recent years, wine consumers have been striving for information (Faria et al., 2020). Additionally, wine consumers are often characterized as risk-averse. Therefore, addressing information asymmetries and avoiding adverse selection is crucial (Gocrekus & Nottebaum, 2011). Electronic marketplaces, like Vivino, can play a pivotal role in providing intrinsic, extrinsic, and personal product attributes for wine and eliminate this asymmetry. However, this is a double-edged sword as this type of market can lead to collusion (Campbell et al.,2005) and information overload (Grover et al., 2006).

2.2.2 Leveraging Machine Learning's Feature Importance

With the abundance of product characteristics in electronic marketplaces, discerning which attributes are most significant to buyers is essential to prevent information overload. Machine learning and statistical models provide a solution, not only by predicting customer satisfaction based on these features but also by understanding the reasoning behind the outcomes (Breiman, 2001b). In today's complex marketplaces, accuracy alone is insufficient. Understanding the "why" behind these predictions is imperative (Doshi-Velez & Kim, 2017). Several methods facilitate model comprehension including feature importance, counterfactual explanations, adversarial training, or influential samples (Bhatt et al., 2020). Among these, feature importance is highly relevant to this study as it helps to assess which product attributes significantly influence the prediction of consumer review ratings. This analysis relies on decision trees as the base estimator for the predictive models, chosen for their high interpretability (Hastie et al., 2009). The relative importance of a feature in a single decision tree can be assessed by measuring the enhancement in the squared error when the features are used to split. The following formula represents the overall improvement contributed by the feature l over J internal nodes in a decision tree T (Breiman & Ihaka, 1984).

$$\tau_l^2(T) = \sum_{t=1}^{J-1} \hat{\imath}_t^2 I(v(t) = l)$$

Equation 1: Calculation of Feature Importance for Feature, equation from: (Breiman & Ihaka, 1984)

2.3 Classification with Ensemble Models

However, to assess the importance of vital customer satisfaction features, a single decision tree is not enough as it does not capture the complex relationship between the product attributes. As a result, it is essential to employ non-parametric machine learning methods. To improve the ability to capture the diversity of datasets in the era of Big Data, ensemble models have gained popularity across numerous industries and sectors in recent years (Natekin & Knoll, 2013). They can be used for regression and classification, in the following paragraph the focus is on classification relying on the most common algorithms are Boosting and Bagging (Hastie et al., 2009).

Ensemble methods integrate multiple weak learners, defined as only slightly more accurate predictors than random guessing, into a meta-model. This technique forms a robust learner and enhances the ability to capture the variation within a dataset. The resulting model effectively improves the predictive accuracy. This meta-model typically utilizes majority voting based on the predictions from the weak learners to make the final classification. Consequently, wrong predictions may be canceled out by aggregating them. In addition, ensemble models capture a more comprehensive view of the overall process, rather than a limited one, because they are composed of many local weak learners (Dietterich, 2000).

However, training these learners using identical datasets can lead to correlated errors, undermining the ensemble model's correctness (Géron, 2023). To counter this, varying the training data, particularly through Bootstrapping (Breiman, 1996b), can be used. Bootstrapping involves drawing random samples with replacements from the original dataset with the same probability of choosing the data (Efron & Tibshirano, 1994). The sampled number of bootstraps is then used to fit a classifier for each sample. This process creates different training datasets and can cause some data points to appear more prevalent (Dietterich, 2000). The so-called Bagging provides the opportunity to benefit from the data, not for the fitting process used as additional validation for the generalization ability of the model, known as out-of-bag evaluation

(Breiman, 1996a). Nevertheless, it is crucial to note that Bagging enhances the meta-model's precision only if small variations in the training set lead to significant changes in classification (Breiman, 1996b). Models like neural networks, decision trees, and logistic regression tend to be unstable and fulfill this criterion (Breiman, 1996c). For instance, Bagging is used in the Random Forest algorithm.

Distinct from Bagging, Boosting involves training models sequentially, creating a dependency on the performance of the preceding model. This sequential training is key to its methodology. After the initial fitting, Boosting adjusts the weights of misclassified instances, thereby accentuating their significance in subsequent predictions. The final meta-model derives its predictions by taking a weighted average of these individual learners (Freund et al., 1990). In this composite prediction, models that demonstrate superior predictive accuracy are assigned greater weight, reflecting their higher reliability (Ridgeway, 1999).

Both methods generate multiple models by using different datasets during training. Bagging utilizes various random samples, whereas Boosting adjusts the training set by weight (Dietterich, 2000). The following graph illustrates this weight adjustment process, distinguished in sequential and parallel.

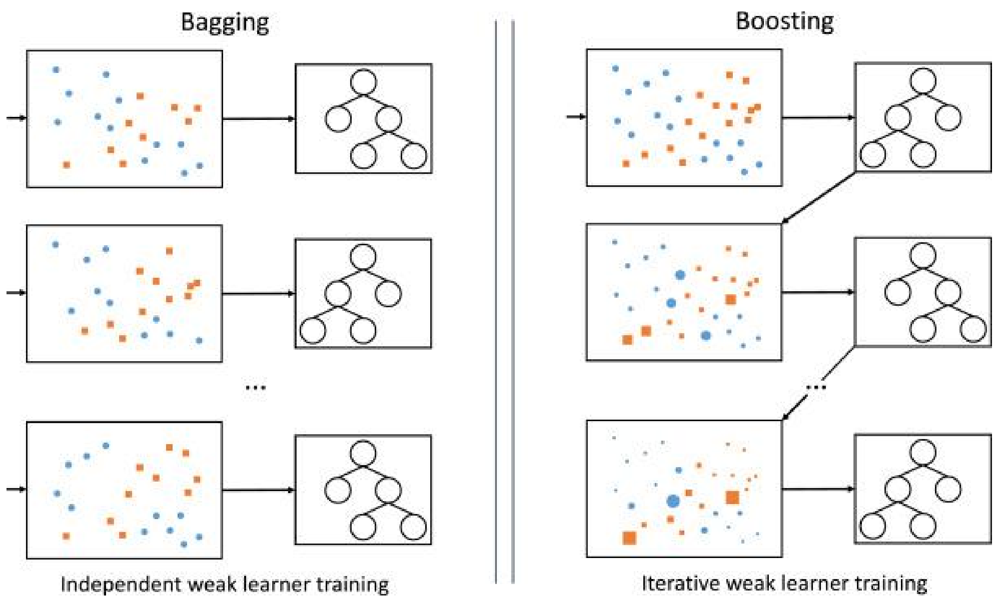


Figure 1: Comparative Illustration of Bagging vs. Boosting in Ensemble Learning, graphic from: (González et al., 2020)

Another emerging advantage of using an ensemble model is the determination of feature importance. The previously introduced equation 1 for determining the importance of a feature in a single decision tree is modified to cumulate the feature's importance across all trees, which is then averaged. This provides additional stability due to the averaging effect (Hastie et al., 2009) and this can indeed help to determine the most significant features for customer satisfaction with a greater level of certainty.

2.3.1 Random Forest

After exploring the foundational concepts of ensemble models, with a focus on Bagging and Boosting, it is pertinent to delve into a specific and highly effective instance of these models: Random Forests. A key issue with single decision trees is their tendency to overfit as the complexity increases, in particular the number of nodes and features (Ho, 1995). This overfitting leads to models that capture noise and are too closely aligned to the training data, resulting in a significant discrepancy between training and test accuracy (Hastie et al., 2009). Consequently, it is important to employ Random Forest, as it addresses this problem by constructing multiple trees. Each tree is trained on a randomly sampled subset of data and considers a random subset of features at each split. At the end of the process, the class is determined by the majority vote of each tree. As a consequence, it is likely to have more correct predictions (Breiman, 2001a). This can also decrease the variance and make it possible to deal with imbalanced data as the algorithm leverages the averaging effect on ensemble trees (Hastie et al., 2009). These characteristics are critical in assessing the impact of specific product attributes on customer satisfaction since the elaborated recommendations should be generalizable by avoiding overfitting.

Moreover, the second crucial characteristic, the randomized features to build the model, allows the capture of complex underlying patterns. This enhances the precision of each tree in the model while decreasing their respective inter-correlation. This provides a multifaceted global picture by the local experts (Breiman, 2001a). This feature is particularly beneficial in the context of this study, where the determinants of customer satisfaction are vast and complex.

Additionally, this algorithm applies a sort of Bagging to resample the training data of the whole datasets facilitating the Out-of-Bag evaluation (Hastie et al., 2009). In Random Forests, the significance of each variable is determined by its contribution to enhancing decision-making at every point where a tree divides into branches. This value is aggregated for each variable across all trees in the forest to gauge its overall importance. This also makes it possible to integrate all meaningful variables in the model (Hastie et al., 2009). In turn, this can be used to assess the individual contribution of product features to customer satisfaction. The following Figure 2 illustrates the predictive process.

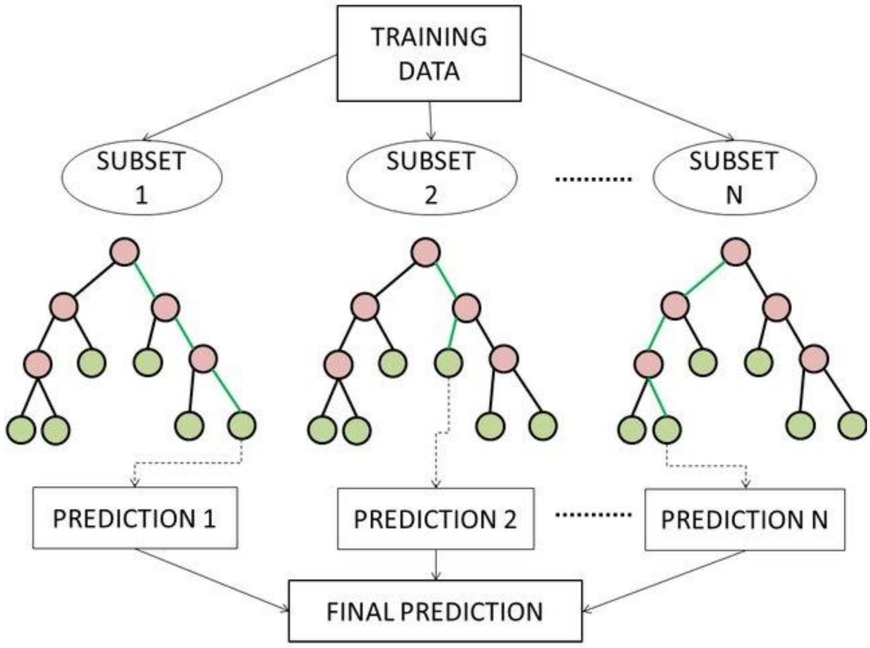


Figure 2: Illustration of Random Forest Ensemble Technique, graphic from: (Laudato, et al., 2020)

2.3.2 AdaBoost

Another algorithm that relies on the ensemble method is AdaBoost. Unlike Random Forest, which can be parallelized (Breiman, 2001a), it operates through a sequential learning process. It involves iterative adaptation, initially using a random sample from the entire dataset and continuously employing this data throughout the process. After each fitting process, weight modifications are made, placing more emphasis on the misclassified instances. This approach relies on a simple model that makes marginally more accurate predictions than random chance (Freund et al., 1990). By varying the weights, AdaBoost exposes the model to different data distributions, thereby improving its generalization capabilities (Hastie et al., 2009). This is one of the key characteristics that was leveraged in this study.

For a final prediction, each fitted learner contributes to a linear, probabilistic function rather than relying on majority voting as in Random Forest (Drucker et al., 1993). Originally, AdaBoost was limited to binary classification, but it was extended to multi-class scenarios through the SAMME algorithm (Hastie et al., 2009). This algorithm adjusts the weight update function with a scaling log parameter, enhancing the performance of weak learners in multi-class settings. The following graph illustrates how each weak classifier contributes to the overall classification in the meta-model.

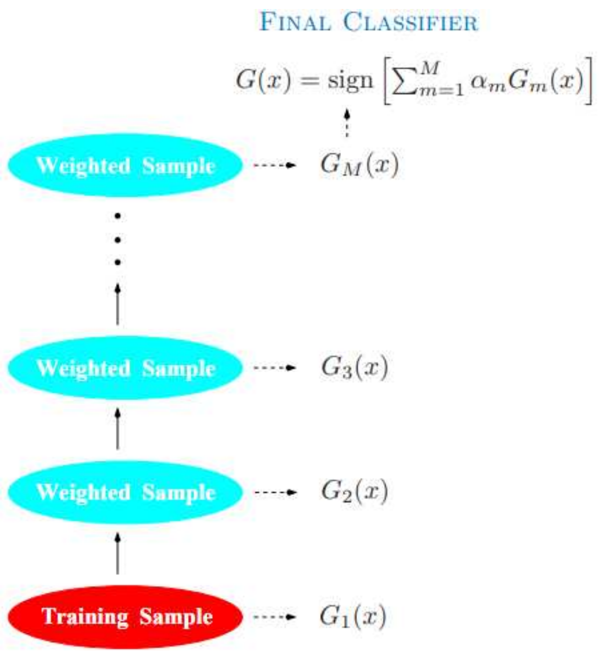


Figure 3: Illustration of AdaBoost Ensemble Technique, graphic from: (Hastie et al., 2009)

In the context of this analysis, AdaBoost is applied using decision trees as the base learner, allowing for the determination of feature importance. The significance of each feature is calculated as an average across all fitted models. Due to the averaging effect, AdaBoost tends to be steadier than single decision trees. However, AdaBoost might fail to take into account all features due to the nature of the Boosting algorithm (Hastie et al., 2009).

2.3.3 Gradient Boosting

Besides AdaBoost, Gradient Boosting relies on Boosting. This algorithm merges machine learning and mathematical optimization (Friedman, 2001). Like AdaBoost, Gradient Boosting operates sequentially to predict classes. However, unlike AdaBoost, the algorithm does not solely rely on adjusting the dataset through weight updates. Instead, Gradient Boosting concentrates on reducing the loss function, which is the discrepancy between the actual and predicted value, of the entire ensemble model. To achieve its objective, the algorithm seeks to rectify the inaccuracies of its predecessor used to classify the model. The recently fitted estimator exhibits a strong correlation with this gradient of loss (Natekin & Knoll, 2013). However, the sequential approach makes the model prone to overfitting and losing its ability to generalize due to the ease of adding more trees to the ensemble. Thus, a range of regularization should be considered. These include fitting the model with a randomly drawn subset of the training data at each iteration and each tree, thereby reducing preprocessing time (Sutton, 2005). Another approach is to utilize a shrinkage parameter, which applies proportional penalization to each subsequent fitting process and slows down the learning process. This can avoid too complex models (Friedman, 2001). Alternatively, trees can be fitted to a random subset of features, referred to as Stochastic Gradient Boosting (Friedman, 2002). The following Figure represents the sequential process of this ensemble method in the predictive task.

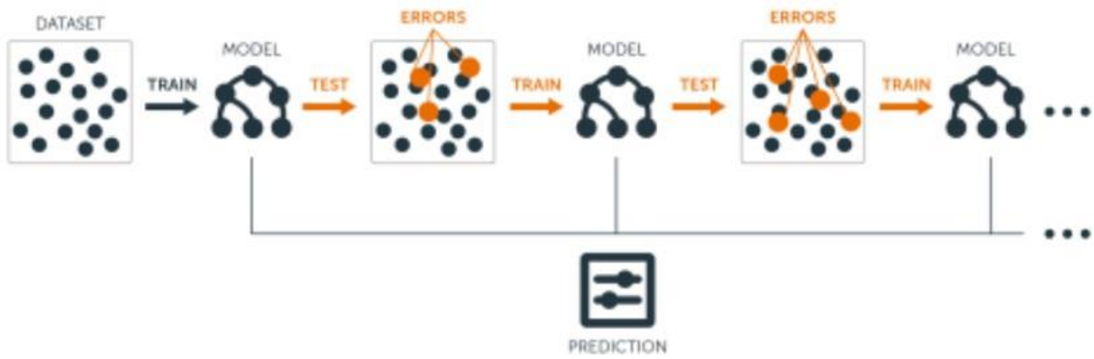


Figure 4: Illustration of Gradient Boosting Ensemble Technique, graphic from: (Boehmke & Greenwell, 2019)

Overall, Gradient Boosting is a versatile algorithm capable of catering to specific requirements and processing large volumes of data. Hence, it is reasonable to deploy this algorithm for deducing the significance of various features in forecasting customer satisfaction. In terms of feature importance, this algorithm operates similarly to AdaBoost by utilizing the cumulative effect of sequential trees (Hastie et al., 2009).

2.3.4 XGBoost

A further refinement of Gradient Boosting is XGBoost. This algorithm stands out in the field of Machine Learning for its exceptional performance in various applications. It is an improved version of the previously discussed traditional Gradient Boosting, with a key distinction in its objective function. This function is designed to minimize the loss and to modify the original one by including an additional regularization. This regularization term imposes a penalty on overly complex trees and thereby promotes better generalization ability with simpler trees. As a result, the efficiency is augmented. If the value of the regularization term equals zero, XGBoost operates as a normal gradient Boosting tree (Chen & Guestrin, 2016).

The modified loss function is depicted in the following equation. Its difference is represented by the **Ω** in the modified loss function, where l captures the loss between the predicted and actual value of the target.

$$L(\phi) = \sum_{\mathrm{i}} \mathrm{l}(\hat{y_\imath}, \mathrm{yi}) + \mathrm{k}\Omega(\mathrm{f_k})$$

Equation 2: XGBoost Regularized Loss Function Formulation, equation from: (Chen & Guestrin, 2016)

In addition, this approach enhances parallelization capabilities compared to previous models (Johnson & Zhang, 2014). Besides the algorithmic improvements, XGBoost is also renowned for its speed and scalability, making it a preferred choice for handling large datasets. It efficiently manages sparse data, can be run in parallel or distributed environments, and supports out-of-core computation.

These features contribute to its ability to swiftly process vast amounts of data. Lastly, XGBoost incorporates advanced techniques to optimize tree construction, further refining its ability to find the best solutions in terms of impurity reduction (Chen & Guestrin, 2016). As with the other Boosting tree algorithms, the feature importance grows as it is more frequently used for critical decisions (Hastie et al., 2009).

3 Data

After providing the background for the study's key components and their theoretical match, this chapter describes the methods used for collecting and processing data to assess key features of Portuguese wine. It focuses on web-scraping from Vivino, detailing the pre-processing steps and categorizing product features. The application of topic modeling to incorporate user comments is also discussed. An exploratory data analysis ensures the data's representativeness for making generalizable recommendations. The complete dataset and code are available on GitHub.

3.1 Data Extraction

To determine the most important product attributes of Portuguese wines through predictive analysis, gathering comprehensive data was crucial. This dataset includes the current customer satisfaction measured by consumer review ratings, ranging from 1.0 to 5.0 (Westbrook, 1980). In addition, it encompasses the intrinsic (Akdeniz et al., 2013), extrinsic (Symmank, 2018), and personal (Kim et al., 2014) product attributes previously defined. Notably, the predictive modeling process concentrated on specific consumer review ratings, as illustrated in Figure 5 and referred to as 'review_rating'. This approach aims to garner deeper consumer insights rather than depending on the broad overall rating for a wine's vintage, illustrated in Figure 6 and further named as 'ratings_average' which was very present on the website's interface.

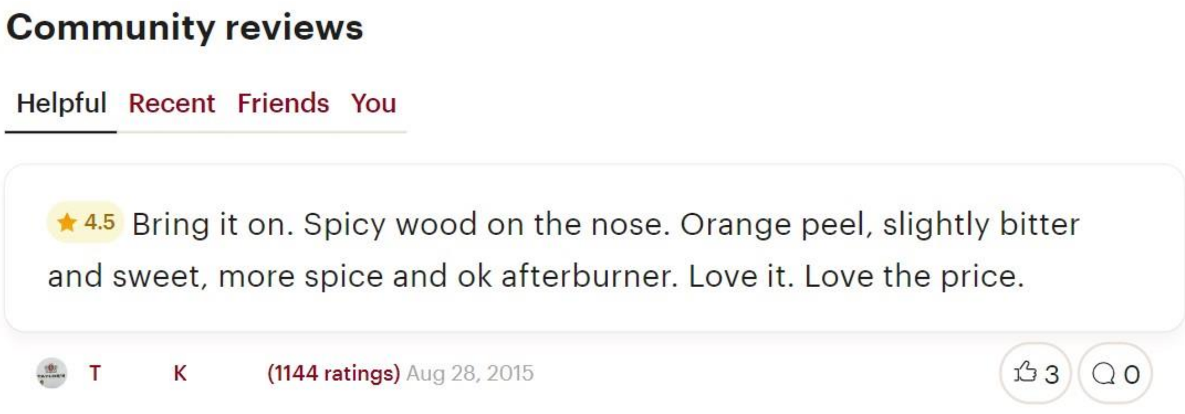


Figure 5: Vivino Consumer 'review_rating' Example, graphic from: (Vivino, 2023b)

Figure 6: Vivino 'ratings_average' for Wine Vintage Example, graphic from: (Vivino, 2023b)

For this purpose, the data was meticulously extracted from Vivino, utilizing an API and a comprehensive three-step web-scraping process. This method converted the website's product attributes into a structured format conducive to machine reading. Initially, the process involved setting specific parameters to retrieve Portuguese wines and their product attributes. Since the Portuguese wine prices were unavailable, they were substituted with German prices to ensure external verification with other websites. The price range was set from 0 to 500 Euros to capture the diverse price spectrum of Portuguese wines. Additionally, the requests were arranged in ascending order. This setup enabled reproducibility and facilitated external and internal verification of data accuracy through random sampling of the output and comparison with the website. The API configuration enabled data collection of 25 wines per request. The number of requests was set to 100, as the amount of unique wines beyond this scope varied insignificantly. As a result, the dataset comprised 2,500 wines with 34 product attributes, including intrinsic and extrinsic attributes associated directly or indirectly with the wine. Since alcohol level data was not accessible via this API, a separate process using a URL request was employed to convert JavaScript into HTML for this specific attribute.

In the final step, the 'wine_id' and 'vintage' identifiers from the web-scraped feature database were used to extract the comments via a URL request. This iterative data collection process retrieved the target variable 'review_rating', along with comment-specific attributes and comments in English. To overcome frequent connection interruptions, this process was limited to 20,000 seconds. Furthermore, comments beyond the first 100 pages were excluded to maintain the relevance and representativeness of current customer satisfaction. As a result, the final review dataset comprised 464,109 'review_rating' observations and 19 product attributes, predominantly personal.

3.2 Data Preprocessing

The previous chapter outlined the methodology for data collection and extraction, resulting in the creation of two comprehensive datasets. As this data obtained from web scraping is in raw format, it is crucial to preprocess it to secure usability. To ensure the representativeness of the extracted product attributes based on extrinsic, intrinsic, and personal attributes, the data compilations were preprocessed separately. On the one hand, the feature dataset mainly covers the extrinsic and intrinsic product attributes. On the other hand, the review dataset covers personal attributes. In alignment with ethical standards and privacy protection, any personal identifiers, such as user names, were diligently removed during preprocessing.

Initially, a new primary key was created by combining the 'wine_id' and 'vintage' fields in the feature dataset, and the 'wine_id' and 'review_referring_vintage' fields in the review dataset. This was crucial as the 'wine_id' remained constant across different vintages, despite differences in attributes such as taste or flavor for each vintage. Thus, this new primary key strategy was instrumental in preventing any loss of information during the merging process. Subsequently, all duplicate 'review_id' entries were promptly eliminated from the review dataset. This step was essential, as the comment section for each wine was defined by the 'wine_id', regardless of the vintage. Consequently, the original dataset contained numerous duplicates, since the comments for each wine encompassed all its vintages. To preserve the unique characteristics of each wine type, a nuanced approach was adopted for handling missing values, rather than immediately deleting them. For instance, fizziness, typically reported only for sparkling wines, and tannin, reported only for red wines, were key considerations. Missing values for these attributes in other wine types, such as white wine, fortified wine, and rosé wine, were replaced with zero values. This step was crucial to ensure the specific characteristics of the respective product attributes were accurately represented. Subsequently, for other product features with missing values, we imputed the data by replacing categorical values with the mode and numerical values with the median, thereby maintaining the original distribution of the dataset (Saar-Tsechansky & Provost, 2007). Additionally, two indicators were created to ascertain whether there were interesting facts or a wine description. In both datasets, any observations where the threshold of missing values fell below 1% were dropped to ensure data quality. In the review dataset, the languages of the users were combined to consolidate similar language backgrounds, such as merging Br-PT and Pt-PT into PT, facilitating a more unified analysis.

After merging the two datasets using the primary key, a final step of preprocessing was performed to generate more easily interpretable predictions. The continuous individual consumer 'review_rating' for each review was categorized into the nearest 0.5 rating interval. This alteration transformed the prediction task from a regression to a classification task, yielding more fitting predictions. As a result, the predicted values are constrained to a range of 1 to 5 and cannot be negative. Additionally, no decimal values beyond 0.5 will be generated.

Finally, the extraneous attributes that were only included in the dataset for technical reasons were removed from the reviews (e.g., SEO name). The final dataset for modeling is comprised of 80,021 observations along with 73 product attributes. These were allocated, as defined in chapter 2.2, in their respective product attribute categories, as represented below.

Product Attribute Category	Product Attribute
Intrinsic Information	‘wine_style’, ‘wine_style_description’, ‘grapes’, ‘body_description’, ‘acidity’, ‘sweetness’, ‘intensity’, ‘tannin’, ‘fizziness’, ‘alcohol’, ‘amount_grapes’, ‘wine_type’, ‘vegetal’, ‘oak’, ‘spices’, ‘tropical_fruit’, ‘non_oak’, ‘earth’, ‘tree_fruit’, ‘floral’, ‘black_fruit’, ‘microbio’, ‘dried_fruit’, ‘citrus_fruit’, ‘red_fruit’, ‘amount_flavor_words’
Extrinsic Information	‘wine’, ‘winery’, ‘region’, ‘vintage’, ‘ratings_average’, ‘ratings_number’, ‘all_vintage_ratings_average’, ‘all_vintage_ratings_count’, ‘price’, ‘price_type’, ‘interesting_facts’, ‘bottle_type’, ‘has_wine_style_description’, ‘has_interesting_facts’, ‘beef’, ‘mild and soft cheese’, ‘blue cheese’, ‘shellfish’, ‘sweet desserts’, ‘veal’, ‘lean fish’, ‘appetizers and snacks’, ‘game (deer & venison)’, ‘vegetarian’, ‘cured meat’, ‘mature and hard cheese’, ‘aperitif’, ‘lamb’, ‘fruity desserts’, ‘pasta’, ‘pork’, ‘rich fish (salmon & tuna etc.)’, ‘poultry’, ‘vegetal’, ‘amount_food_words’
Personal Information	‘review_rating’, ‘review’, ‘likes_on_review’, ‘comments_on_review’ ‘user_language’, ‘user_follower’, ‘user_following’, ‘user_total_given_ratings’, ‘user_total_sum_ratings’, ‘user_total_given_reviews’, ‘user_purchase_made’, ‘ review_rating_label’

Table 2: Product Attribute Categorization for Gathered Wine Data

3.3 Topic Modeling

After completing the extensive data preprocessing outlined in the previous chapter, a challenge arises in how to incorporate the 80,021 comments that were associated with a specific consumer rating review. The study aims to integrate personal experiences expressed in the comments, which are unique and comprise 30,979 words. It is not feasible to incorporate each comment as a separate feature. Therefore, it is essential to decrease this high dimensionality by utilizing statistical techniques that summarize broader content instead of concentrating on individual words and comments. This approach will facilitate the identification of keywords or themes that encapsulate the overall personal experience.

One approach to achieve this objective is the implementation of the tf-idf scheme, short-for-term frequency-inverse document frequency. This method determines how often a word appears in a given comment and relates it to the inverse frequency across all comments, resulting in a higher weight for more unique words (Salton & McGill, 1983). However, it does almost not account for any statistical connections between or within comments, as it exclusively concentrates on word frequency (Blei et al., 2002). An effective method to overcome this issue is Latent Semantic Indexing (LSI). LSI follows the principle that similar concepts tend to appear in similar documents, such as comments in this case. It calculates the frequency of unique words across the comments to evaluate the relationship between words and their respective distribution. After constructing a matrix with this information, LSI applies Singular Value Decomposition (SVD) to reduce dimensionality using cosine similarity. LSI distills the original matrix into a format that emphasizes the underlying patterns and connections among words (Deerwester et al., 1990).

Although LSI is a promising approach, it does not capture certain topics of the comments to reduce the dimensionality of the given reviews. To achieve this, Latent Dirichlet Allocation (LDA) is utilized (Blei et al., 2002). LDA allows for the detection of the common content of comments, facilitating the distillation of a large volume of comments into broader summaries. By aggregating all observed comments into a corpus - a collection acting as a dictionary based on the words used in the comments - it becomes possible to discern content not directly observable (Mimno & McCallum, 2007). For this purpose, LDA employs Bayesian Inference to detect latent topics in this collection of documents. Furthermore, this allocation algorithm assumes that each comment is composed of a small number of latent topics and each word within this comment is generated by a particular topic. The primary challenge is that only the words in the comments are observable, whereas the topics and their distribution within each comment are not. The left part of Figure 7 depicts this challenge, showing the observable word w, due that depicted as gray, within the comment N in the corpus M. The hidden or unknown components, namely the distribution of topics within a comment θ, the assignment of topics to words in a comment Z, and Dirichlet priors α on the per-comment topic distributions and β on the per-topic word distribution, affect the number of topics each comment covers and the number of words each topic is likely to cover, respectively. As these elements are unobservable, they need to be approximated to the complex posterior distributions that are difficult to compute directly in LDA. Consequently, a variational inference and expectation-maximization algorithm is used to iteratively improve the approximation of these distributions, leading to a more accurate model of the underlying topics (Blei et al., 2002). The right part of Figure demonstrates this process, representing the approximation by γ and Φ.

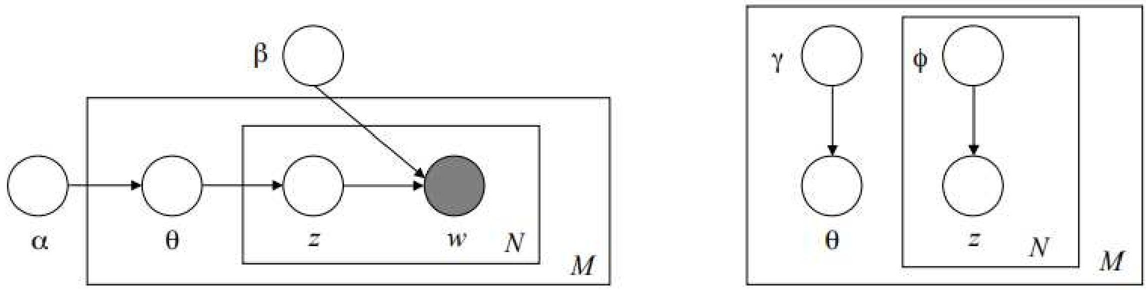


Figure 7: Graphical Comparison of Original LDA vs. LDA Approximation Method, graphic from: (Blei et al., 2002)

As LDA is an unsupervised learning algorithm, it may lack context that is interpretable to humans. Additionally, determining the model's accuracy can be challenging (Chang et al., 2009). To overcome these issues, a frequently utilized method is perplexity, which verifies correctness by assessing the model's performance on unseen data. However, this does not indicate how the outcome compares to human perception. Therefore, this metric is not utilized to ascertain the ideal number of topics for capturing the most important personal experience. Instead, the quantitative method OC-Auto-NPMI is employed (Lau et al., 2014). This technique utilizes an automated approach with Wikipedia data to mimic human interpretability (Newman et al., 2010) through the inclusion of Normalized Pointwise Mutual Information.

Consequently, this metric was used to identify the optimal number of topics and reduce the dimensionality of comments. To enable the deployment of the LDA model, the observed text data was cleaned by removing irrelevant characters, symbols, numerical terms, and emoticons. Besides that, all stop words were eliminated, as such terms do not hold any significant meanings for the subject. The list of English stop words was unmodified, contrary to some recommended practices, to capture a broader understanding of the comments. Split and tokenized words were reverted to their original stems, and words with low frequency were eliminated. As a result, a reliable dictionary was created. This dictionary was also capable of recognizing common bigrams, such as 'Vinho Verde'. Finally, the LDA model was tuned to the number of topics through different document-topic α and word-topic β parameters to enable the variational inference algorithm introduced previously. Coherence, measured by OC-Auto-NPMI, was then calculated for all permutations. The subsequent graph illustrates the average coherence for each topic using different parameters. The most optimal model shows a coherence of 0.665.

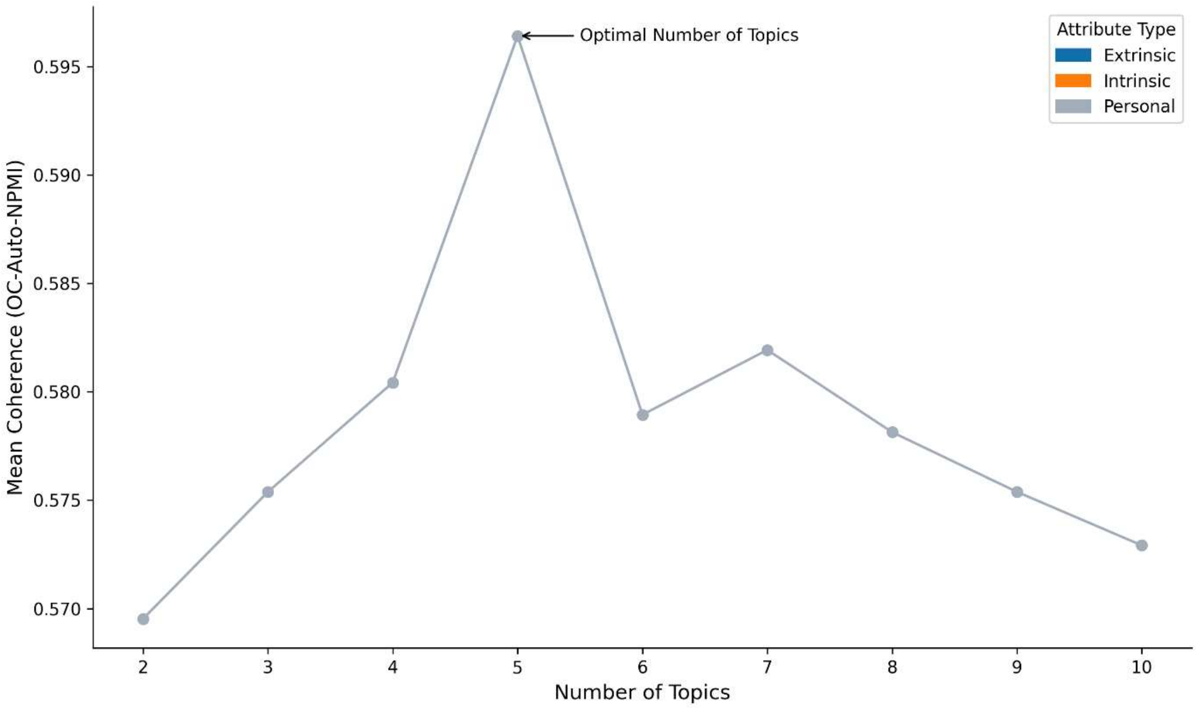
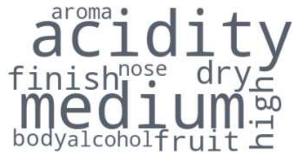


Figure 8: Coherence Score Across Different Topic Numbers for LDA Model

The graph indicates that coherence declines after the average peak for five topics. Thus, reducing the comments to this number appears appropriate. Additionally, this conclusion is supported by the decreasing trend for additional topics. However, it should be noted that a higher number of topics may result in better coherence, but this was beyond the scope of this analysis. As a result, the reviews were reduced to these five topics. The discovered topics can be summarized manually by relying on the top ten significant words: affordable, sweet, analytical, robust, and vintage.

Topic 1: Affordable

Topic 2: Sweet



Topic 3: Analytical

Topic 4: Robust

Topic 5: Vintage

Table 3: Top Words Representing the Five Main Topics Derived from Wine Reviews

3.4 Descriptive and Exploratory Data Analysis

After gathering and preprocessing the data, it is essential to evaluate its representativeness and its suitability for making generalizable recommendations. The dataset comprises 80,021 observations and 73 product attributes, categorized into 35 extrinsic, and 26 intrinsic, 12 personal characteristics. The 'review_rating' data range from April 28, 2012, to October 6, 2023. Figure 9 depicts the increasing number of observations per year until 2022, providing a comprehensive representation for actual feedback. The drop observed in 2023 is likely caused by extraction during the year. The data offers a robust foundation for understanding key factors behind recent consumer satisfaction, serving as a reliable source of actual feedback. The increasing trend for consumer ratings can also be interpreted as rising demand for Portuguese wine, which would prove the forecast of the growing production of Portuguese for 2023 (International Organisation of Vine and Wine, 2023). Furthermore, the graph illustrated over the years shows a clear indication of preferences, dominated by fortified wines, followed by red wines. In contrast, sparkling wines consistently have the lowest number of observations.

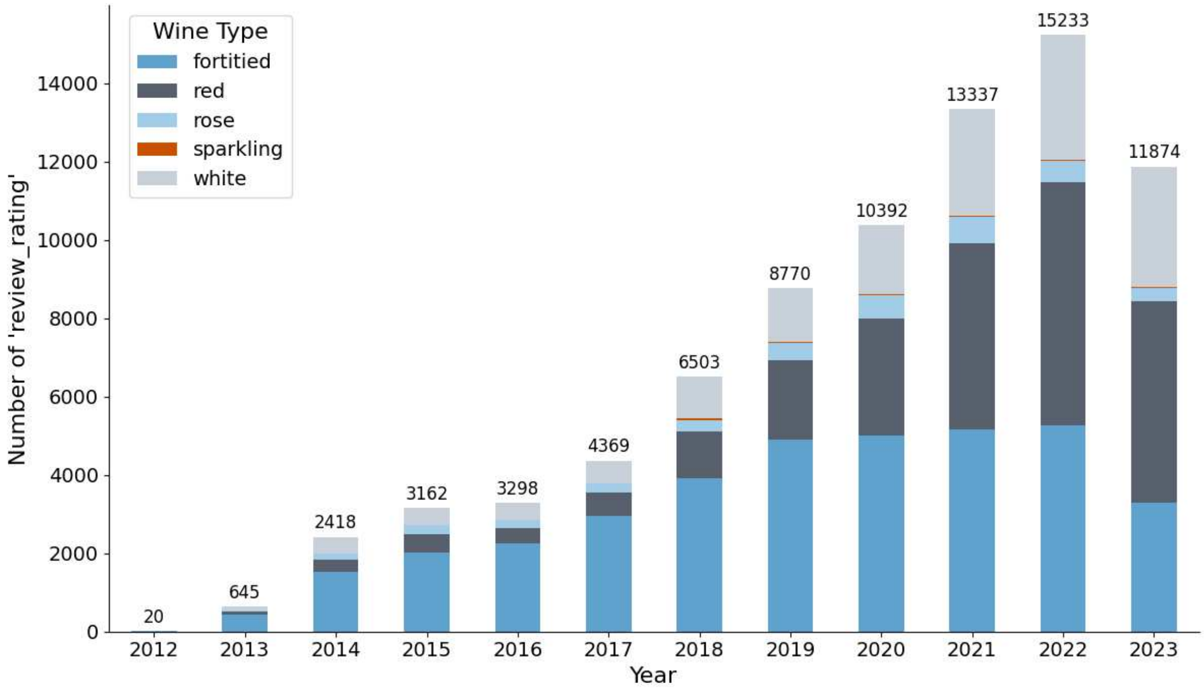


Figure 9: Number of consumer review ratings (2012-2023)

The dataset comprises 17 categorical product attributes, such as ‘winery’, ‘region’, and ‘wines’, each with a high number of labels. This reflects the diverse distribution across the country, influenced by the historical contributions of various cultural groups. Specifically, the data compilation includes specific regions, enabling distinction between quite specific areas such as Alentejo and Alentejano. This aspect is notable as it highlights discrepancies with the officially recognized Denominação de Origem Controlada and Indicação de Proveniência Regulamentada regions. Another observation that emphasizes the historical influence on Portuguese wine is the presence of 23 unique grape varieties. This diversity highlights the uniqueness of Portuguese wines, which is further underscored by the dominance of fortified wines, mainly Port and Madeira. This obesity influences factors such as ‘wine_style’, ‘wine_style’, grape compositions, regional attributes, and non-vintage categorization.

Figure 10 supports the claim and also reveals a history marked by long trade relations with England. The data extraction targeted English comments, and the primary ‘user_language’ identified was English, suggesting that a significant proportion of users are from Englishspeaking countries. This indicates the ongoing demand from the history of fortified wine among English speakers.

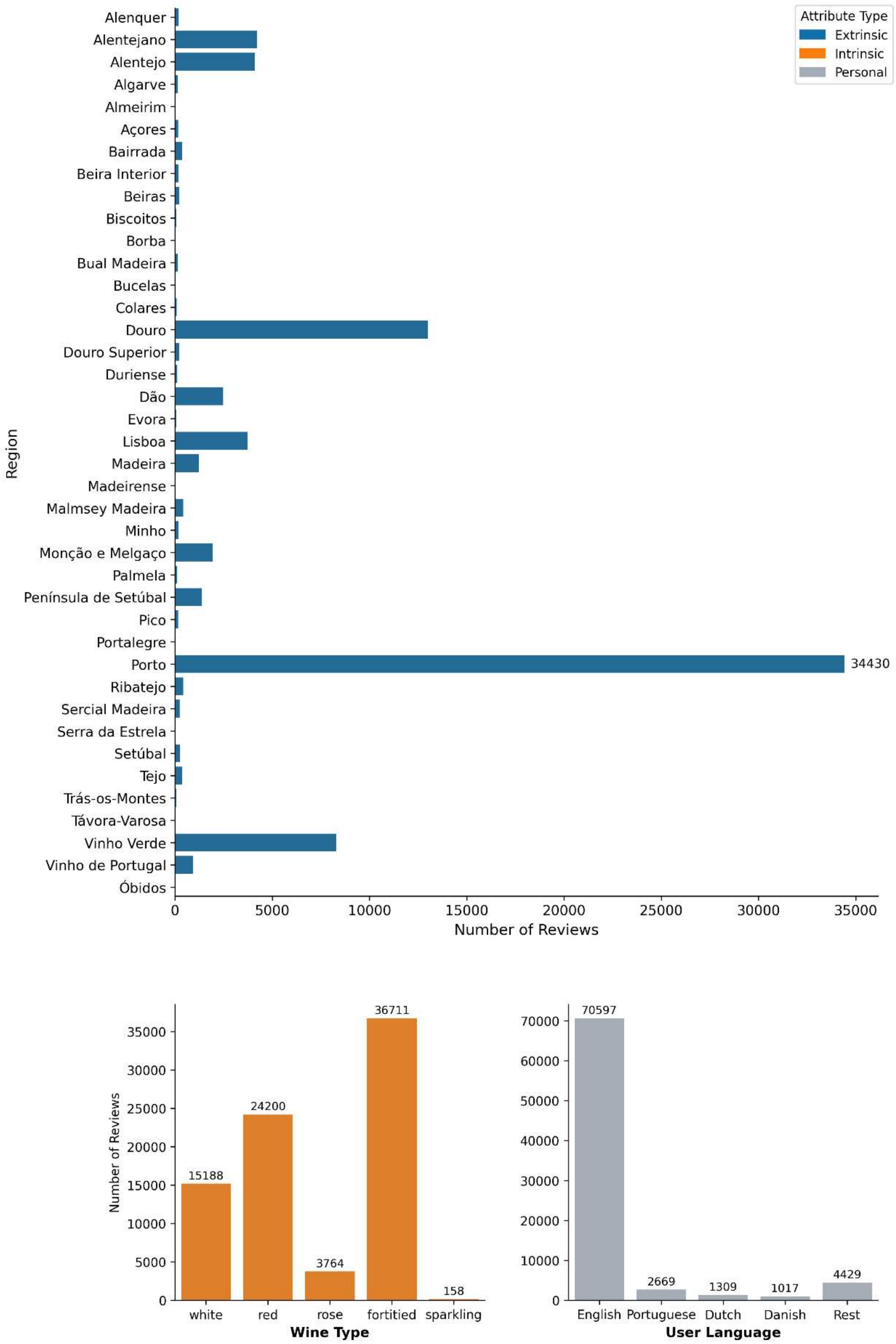


Figure 10: Distribution of Region, Wine types, and Comment Topics in the Wine Dataset

Regarding representativeness, this dominance could also affect the numerical intrinsic product attributes for sensory perception, such as ‘sweetness,’ ‘intensity,’ or ‘acidity’. Nevertheless, upon comparing the data to the original feature dataset distribution, the web-scraped ‘review_rating’ database shows only moderate deviation. When selecting models for predictive analysis, it is important to carefully consider the presence of skewed values in ‘ratings_average’, ‘price’, ‘acidity’, ‘intensity’, and the number of flavor descriptors, as well as outliers in ‘ratings_number’ and ‘all_vintage_ratings_count’. Therefore, this analysis utilized Random Forest, AdaBoost, Gradient Boosting, and XGBoost, which are capable of handling these issues. Table 4 displays the distribution of all numerical values in the dataset, indicating the presence of skewness and heavy outliers.

feature	product attribute	distinct	mean	std	min	median	max	IQR
acidity	intrinsic	1405.0	3.2	0.4	2.1	3.0	4.6	0.3
alcohol	intrinsic	59.0	16.3	3.5	9.0	14.5	21.5	7.0
all_vintage_ratings_average	extrinsic	17.0	3.9	0.2	3.1	3.9	4.7	0.3
all_vintage_ratings_count	extrinsic	1125.0	17184.8	24007.8	25.0	6458.0	87919.0	16090.0
amount_flavor_words	intrinsic	13.0	12.2	1.5	1.0	13.0	13.0	1.0
amount_food_words	extrinsic	5.0	3.5	0.8	2.0	3.0	6.0	1.0
amount_grapes	intrinsic	5.0	3.4	1.3	1.0	3.0	6.0	0.0
comments_on_review	personal	116.0	0.9	5.3	0.0	0.0	228.0	0.0
fizziness	intrinsic	10.0	0.0	0.2	0.0	0.0	4.0	0.0
intensity	intrinsic	1404.0	3.8	0.9	1.5	4.2	5.0	0.8
likes_on_review	personal	346.0	7.2	26.3	0.0	1.0	868.0	2.0
price	extrinsic	1035.0	27.1	39.5	3.5	15.0	470.0	17.8
review_rating	personal	41.0	3.9	0.7	1.0	4.0	5.0	0.7
review_rating_label	personal	9.0	3.9	0.7	1.0	4.0	5.0	0.5
ratings_average	extrinsic	18.0	3.9	0.6	0.0	3.9	4.8	0.3
ratings_number	extrinsic	673.0	10229.4	22301.4	0.0	1811.0	87919.0	5998.0
sweetness	intrinsic	1399.0	3.1	1.4	0.0	2.2	5.0	2.7
tannin	intrinsic	690.0	1.0	1.5	0.0	0.0	4.1	2.8
user_follower	personal	1514.0	357.2	2051.4	0.0	7.0	54688.0	58.0
user_following	personal	1152.0	131.0	488.3	0.0	7.0	5111.0	49.0
user_purchase_made	personal	139.0	2.5	12.1	0.0	0.0	449.0	0.0
user_total_given_ratings	personal	3032.0	865.0	1624.0	0.0	231.0	15236.0	897.0
user_total_given_reviews	personal	2681.0	748.0	1535.7	0.0	157.0	15234.0	715.0
user_total_sum_ratings	personal	19224.0	3224.9	6069.3	0.0	870.8	55669.0	3314.1

Table 4: Distribution of Numerical Values in the Wine Dataset

As stated in the preprocessing chapter, the individual consumer ratings were grouped into bins. Figure 11 illustrates this process and justifies the decision to maintain the original distribution. However, it is important to note that the low consumer ratings in both distributions are highly underrepresented. To ensure the representativeness of the data, the distributions were compared to the original feature dataset once again, and the deviation was found to be minimal.

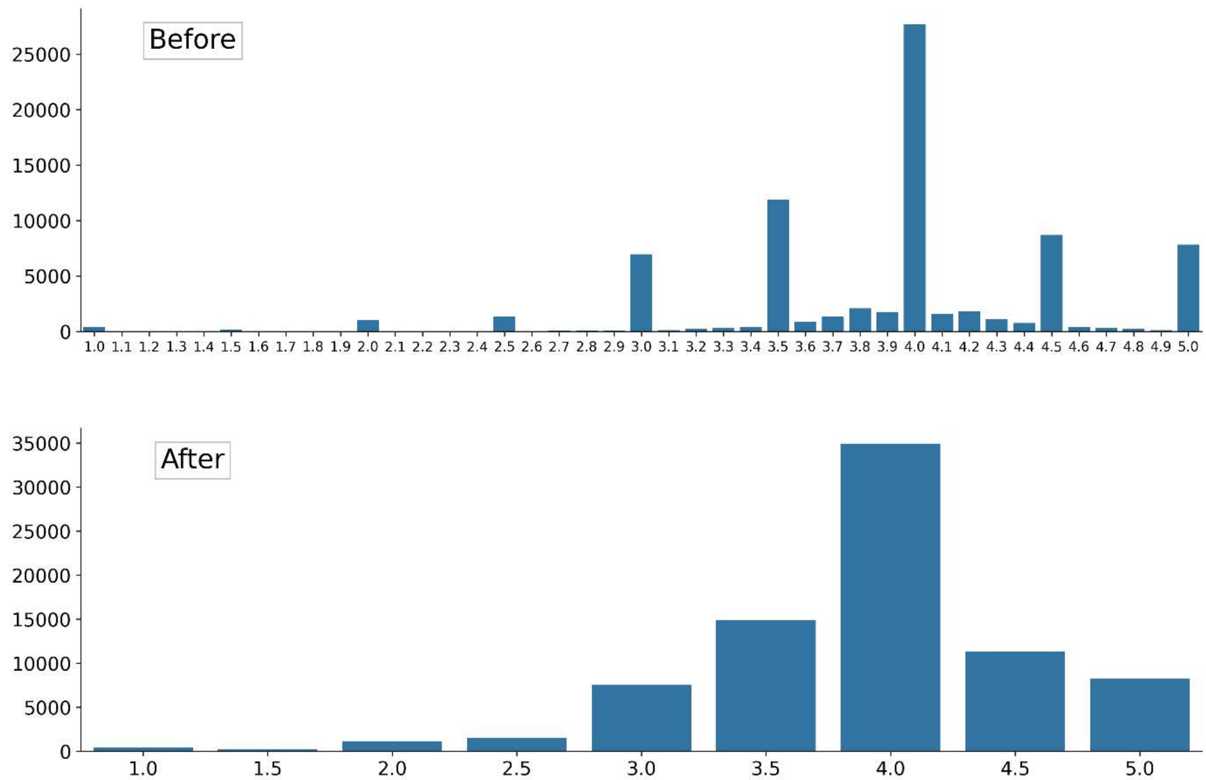


Figure 11: Distribution Comparison for 'review_rating' before and after classification

In summary, despite certain overweight, this exploratory data analysis validates the representativeness of the web-scraped data. Consequently, the dataset can be used for the predictive modeling process and for drawing business recommendations.

4 Results

Building on the groundwork laid in the data collection and preprocessing chapter, this chapter delves into the predictive analytics components. The analysis involves the use of four distinct machine learning algorithms to predict consumer review ratings in a classification task. The results are then used to determine which features were important in making this classification to derive business implications. Finally, this text provides a set of strategic recommendations for applying these implications in the real business world.

4.1 Modeling

The main objective of this analysis is to effectively classify nine distinct classes of 'review_rating' using the previously introduced ensemble models: Random Forest, AdaBoost, Gradient Boosting, and XGBoost. The modeling process has been standardized to ensure a fair comparison between the different models and to gain a comprehensive understanding by utilizing the strengths of each model.

To construct robust and predictive models, duplicates, missing values (Saar-Tsechansky & Provost, 2007), and irrelevant features were removed from the dataset. Additionally, one-hot encoding was used to transform the various labels of categorical data into unique dummy variables to make each feature readable to the machine learning algorithm. The approach improved interpretability and precision in capturing the importance of product attributes while avoiding the numerical bias associated with label encoding (Zelaya, 2019). Next, all features were standardized to a mean of 0 and a standard deviation of 1 to address the outliers and skewed distributions identified in chapter 3.4. This led to more stable model outcomes. Overall, the number of product attributes increased to 257, categorized into 79 intrinsic, 133 extrinsic, and 45 personal. To address the imbalance in lower-valued consumer rating categories, as shown in Figure 11, a training-test split was performed in a 70-30 ratio to ensure sufficient data in both datasets. The dataset was stratified based on the target variable 'review_rating' to ensure more representative training and testing data.

To enhance the model's performance, Bayesian Optimization was used to tune the hyperparameters during the training process rather than using random or grid search, to increase efficiency. This optimization method leverages statistical methods and insights from past evaluations to explore the search space and improve future evaluations compared to grid models. The Optuna API (Optuna, 2023) was implemented to perform this statistical optimization. The hyperparameter spaces for each model are specified in chapter 4.2. These spaces are stored within an objective function and are used to guide the model through predefined areas to aid in its evaluation.

The primary objective of this function was to identify the highest ROC-AUC metric within the given hyperparameter space. This metric was selected due to its suitability for datasets with multiple and imbalanced classes, where traditional accuracy metrics may be misleading. ROC-AUC effectively measures the model's ability to distinguish one class from others, making it ideal for this analysis (Everson & Fieldsend, 2006). In addition, the metric played a crucial role in the five-fold cross-validation process, which was always stratified based on the target variable to maintain class distribution. The average results obtained from this validation are essential in demonstrating the model's ability to generalize, which is a crucial aspect of the generalizable recommendations (Hastie et al., 2009). Additionally, the model training employed an early stopping method to interrupt the process if no improvement was observed in the hyperparameter space after 10 consecutive iterations (Zhang & Yu, 2005). This was crucial since the trial period for Bayesian Inference was set to 100 iterations, preventing the model from stagnating in non-improving hyperparameter regions and the risk of overfitting (Chen & Guestrin, 2016). Table 5 illustrates the number of processed trials.

In addition, the Optuna API was leveraged to have tunable parameters identified, given its capability to evaluate parameter importance. To enhance the efficiency of model training and align with the stated objectives, the least significant hyperparameters were strategically excluded by relying on this output.

After the optimal hyperparameters were determined, the model was fit and integrated into a tailored function. This function was used to evaluate the training and testing accuracy to gauge the model's generalization ability, as well as the classification matrix to assess its ability to distinguish between classes. Additionally, the macro F1 score and recall were calculated to further understand the model's performance. Ultimately, these scores were used to refine the model. The culmination of these methodologies is reflected in the model performance metrics, as detailed in Table 5, offering insights into the effectiveness of each model.

Model	**Best Test ROC-AUC**	**Training Accuracy**	**Test Accuracy**	**Number of Trials**
Random Forest	75.10%	49.70%	47.16%	19
AdaBoost	65.74%	46.95%	46.53%	15
Gradient Boosting	74.86%	52.58%	47.29%	44
XGBoost	75.20%	53.58%	47.17%	23

Table 5: Comparative Performance Metrics of Ensemble Learning Models

Beyond the quantitative performance metrics, the tailored function enabled to ascertain a feature's contribution towards the classification task. The feature importance was calculated based on the impurity (scikit-learn developers, 2023). These distributions are illustrated in the bar charts below (see Figure 12, 13, 14, 15), differentiating between the overall contribution and distinct attribute contributions of extrinsic, intrinsic, and personal. As a result, this evaluation not only assesses their predictive accuracy but also delves into the business implications drawn from their feature importance analysis.

4.2 Model Evaluation & Business Impact

4.2.1 Random Forest

Beginning with the Random Forest model, initially, a significant overfitting issue was observed when using the model's standard parameters, namely a training accuracy of 100% and a testing accuracy of 44%. This discrepancy, along with a modest ROC-AUC of 67%, necessitated a strategic approach to hyperparameter tuning. The result was a more balanced performance, with training accuracy rising to 50% and testing accuracy to 47%, alongside an improved test ROC-AUC of 75%. Such an evolution in the model's metrics underscores its enhanced capacity for generalizability, making it a more reliable tool for predicting consumer review ratings and analyzing feature importance for business insights.

In optimizing Random Forest the following hyperparameter spaces were set. Firstly, a high number of estimators were chosen to enhance the model's robustness against overfitting and to solidify its capacity to generate reliable recommendations across various scenarios. This increase also allowed the model to capture nuanced and diverse product attributes that influence customer satisfaction by leveraging ensemble methods. Furthermore, this hyperparameter range was set to handle the unbalanced dataset, particularly for low-rated consumer ratings, as emphasized by a small criterion for leaf splitting. The considered range of possible maximum tree depths strategically balances the bias-variance trade-off, preventing overfitting while still creating a complex model to incorporate diverse structures for predicting consumer ratings. This balance should make the model applicable to unseen data. For determining the most important features, the model was given the opportunity to select the features during the splitting period from the log2 base of the total number of features, the square of the number of features, or no limitations on features. This approach enabled the possibility of relying on the truly important features while considering the possible complexity.

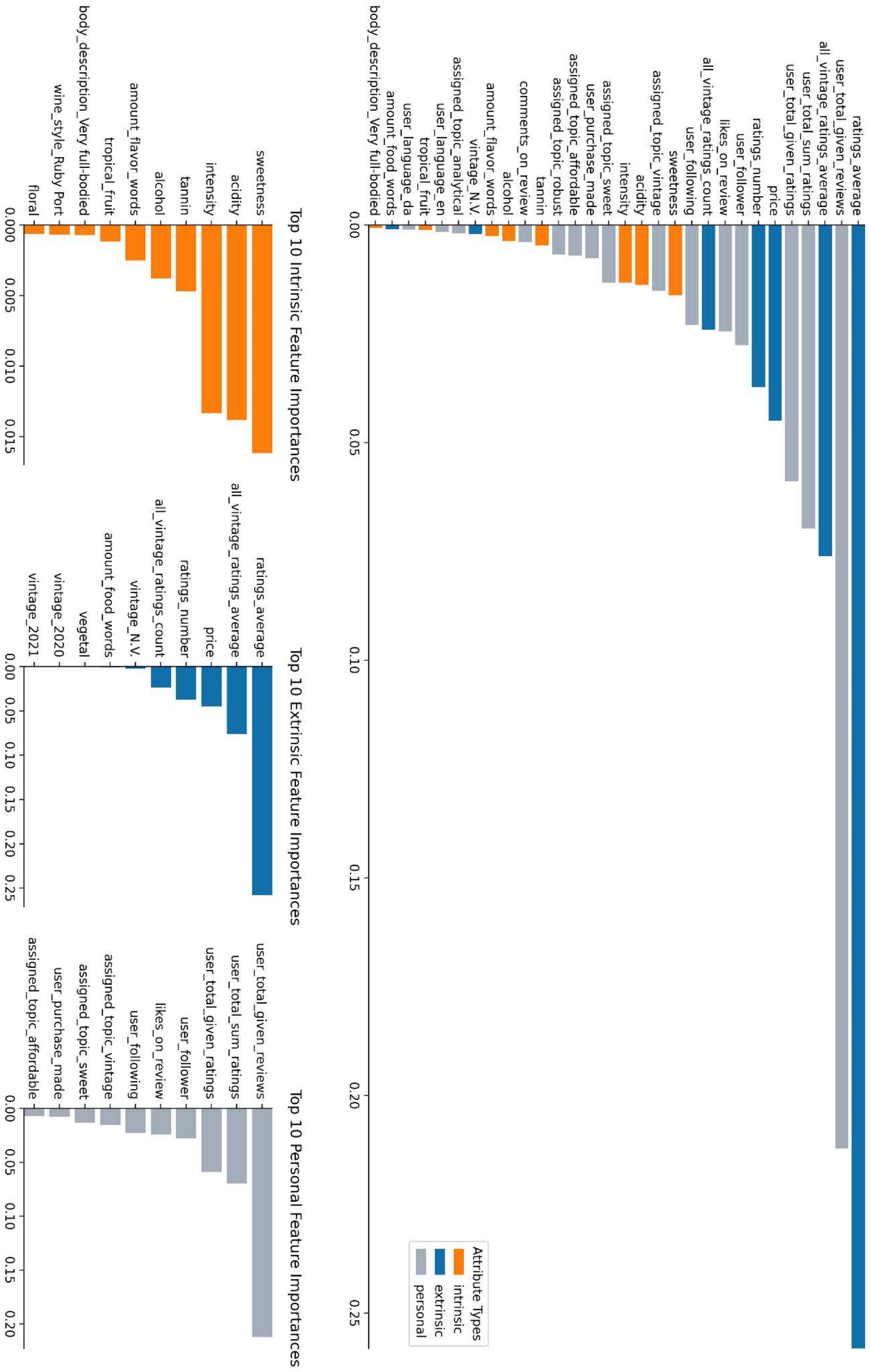


Figure 12: Random Forest Feature Importances by Product Attribute Category

The depicted graph indicates that evaluations made by other consumers significantly impact the prediction of individual consumer wine ratings. The 'ratings_average' emerges as a primary contributor and is supplemented by 'all_vintage_ratings_average', which serves as a substitute when a particular wine vintage lacks enough ratings. The remarkable relevance of 'ratings_number' underlines the importance of the number of ratings a wine receives from other consumers. Thus, it is important to encourage consumers to give a rating to the wine. Furthermore, the value of 'user_total_given_reviews' emphasizes the essential function of user engagement in assessing the individual consumer rating. This is supported by the significance of 'user_total_given_ratings' and 'user_total_sum_ratings'. Therefore, it is crucial to encourage user engagement by writing a comment and rating it.

Shifting the focus to the previously established product attribute categories, the overall importance of intrinsic attributes is low. However, it is evident that sensory perception factors, such as sweetness levels, acidity, intensity, tannin and alcohol content, are more influential in accurately categorizing review ratings than other traits like wine type, grape variety, or wine style description. This is particularly compelling because the dataset is dominated by Port wines. Thus, the focus should be on sensory perception rather than other descriptive characteristics. A trend among consumers for sweet, intense, and acidic wines could be inferred, but this may also be driven by the contribution of the dataset.

In terms of extrinsic features like food pairing suggestions and vintage, descriptors show a lower level of feature importance, whereas price emerges as a relatively more significant factor, reflecting a consumer's cost-benefit advantage. Hence, the price should be more present than other features. This observation is supported by the emerging comment topic affordable, which is part of the top 10 features for personal attributes. For food pairing and taste, it may be beneficial to consider broader categories rather than the current granular ones, as only the number of food pairing words appears as a significant feature.

4.2.2 AdaBoost

Turning now to AdaBoost, we continued our exploration for a robust predictive model. Initially fitting AdaBoost with default parameters yielded an accuracy of 46% for both training and testing, coupled with a ROC-AUC of 66%. This indicates underfitting due to an overly simplistic model that fails to capture the underlying patterns of the product features needed to predict consumer ratings. The feature importance analysis confirms this conclusion. To remedy this situation, hyperparameter tuning was utilized, which yielded improved performance and more reasonable features. The training accuracy was improved by 1 percent and the test accuracy by 1 percent. Furthermore, the ROC-AUC increased to 66%. These results suggest a more reliable model for distinguishing the classes and make it capable of drawing general conclusions about the important features.

To ensure that the model captures the underlying patterns of features affecting consumer satisfaction, the hyperparameter space for the model was defined as follows: The number of estimators in the range was determined to enhance overall classification by creating more nuanced and sophisticated results. The selected number of trees effectively balances the bias-variance trade-off, avoiding overfitting to the training data and poor performance on unseen data while still being complex enough to capture the patterns. Consequently, the model is more generalizable to derive business impacts from the importance of the diverse product attributes. The training process employed a relatively low learning rate to reduce the overemphasis on misclassified observations. As a result, the model's ability to handle class imbalances in low-rated consumer ratings is enhanced. Moreover, the conservative settings of the decision tree's hyperparameters enhance generalization while still enabling the capture of complex relationships. Additionally, this set of base estimators is advantageous for capturing non-linear connections and is essential in estimating feature importance. Furthermore, the implementation of the variation of the SAMME algorithm optimizes the model for multi-class classification tasks. This improves the model's efficiency and expands its usefulness, making it ideal for the consumer rating classification. Having optimized the model's parameters for more accurate predictions, the focus is shifted to understanding which features play pivotal roles, as illustrated in Figure 13.

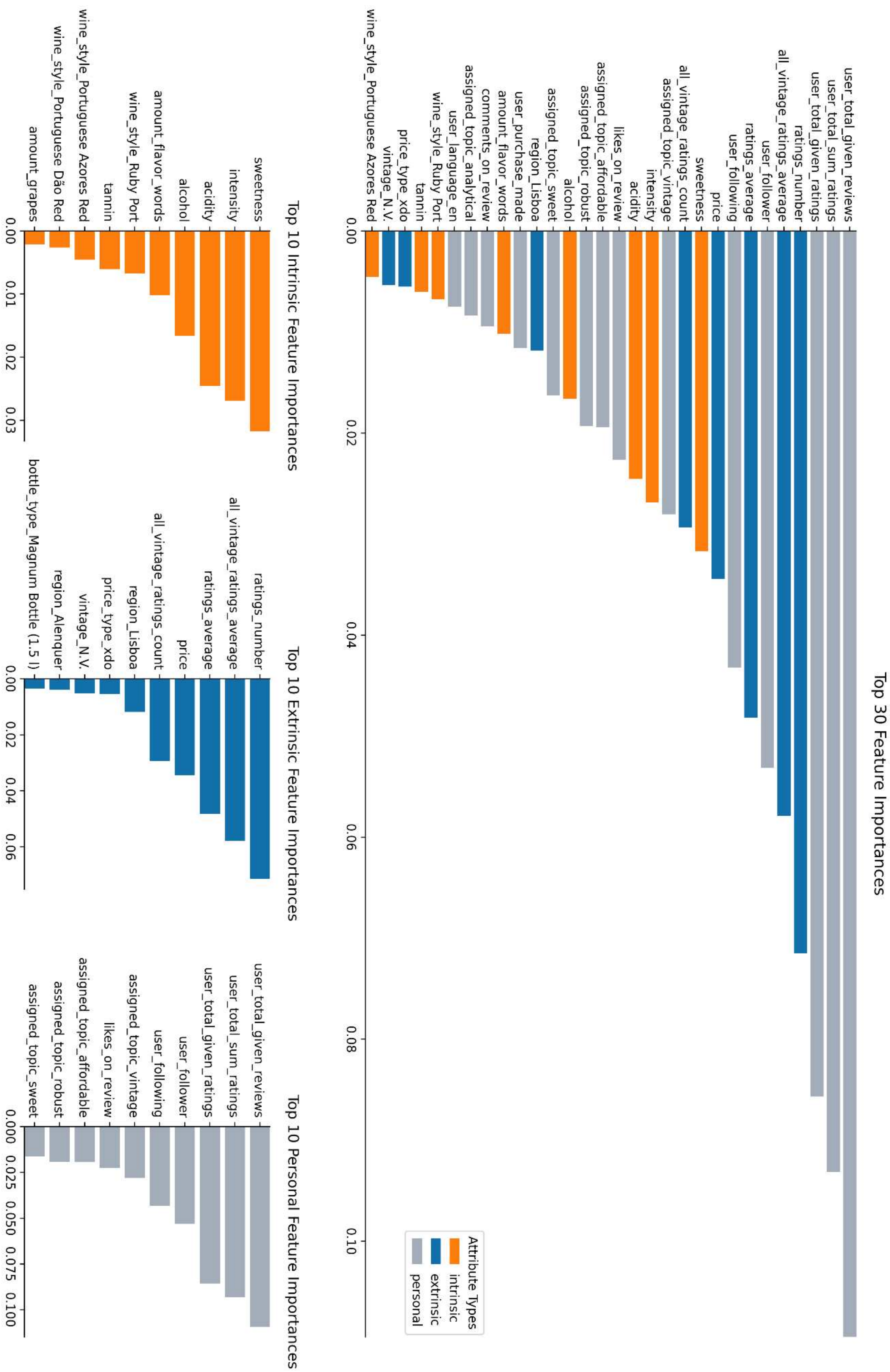


Figure 13: AdaBoost Feature Importances by Product Attribute Category

The graph depicts feature importance for classifying consumer review ratings how significant user engagement is. This is evidenced by the relatively high importance of 'user_total_sum_ratings', 'user_total_given_reviews', and 'user_total_given_ratings'. Consequently, the interaction between users is quite important. Furthermore, the importance of rating frequency is emphasized by both 'ratings_number' and 'all_vintage_ratings_count', underscoring the importance of consistent consumer participation in providing ratings that prioritize frequency.

Upon evaluating the intrinsic attributes of the product, it is evident that sensory perception surpasses all other innate features. This is supported by the substantially high significance of sweetness, acidity, and intensity within this category. In addition, this comprehensive analysis highlights the impact that different wine varietals and vintages can have on the overall experience. Although the importance factor is not particularly high, it is worth mentioning, especially since the labels used for hot encoding, as described in the previous chapter, have high labels. Analyzing the importance of 'wine_style_Portuguese Azores Red', a trend might be observed as it is not related to Port wines.

Moreover, price is a crucial factor in predicting consumer ratings accurately and has significant importance in overall predictions regarding extrinsic product attributes. The significant feature 'affordable', as a personal attribute, further stresses that cost considerations are an integral part of a consumer's perceived value of wine.

4.2.3 Gradient Boosting

Moving further to find the most robust model Gradient Boosting was employed. Fitting this ensemble method with it default parameters resulted in 50% and the testing accuracy was 47%. The ROC AUC was 74%. To improve this, hyperparameter tuning was performed resulting in both accuracies increasing to 52% for the training accuracy and 47% for test accuracy. The ROC-AUC increased to 75%, indicating a more reliable model.

To reach the performance increase and enhance the reliability of feature importance for generating business insights, specific hyperparameters were carefully adjusted. This tuning process aims to balance model complexity with generalization capability. The range for the number of estimators was carefully selected to allow for the detection of complex patterns in consumer rating predictions while avoiding overfitting and loss of generalization. A higher number of weak learners facilitates the identification of intricate relationships for accurate consumer rating classification. Meanwhile, the gradual range for learning rates should ensure the model's ability to generalize the prediction and the feature importance. Additionally, this enables the model to successfully manage imbalanced classes, in this case, particularly low-rated consumer ratings. Moreover, the model continually tracks the presence of this class, resulting in a more resilient model. Subsampling continuously introduces new data points through bootstrapping, also known as stochastic gradient, to reinforce generalizability. Doing so prevents the model from becoming too familiar with the data and reduces variance, resulting in feature importance that is generalizable.

In addition, the hyperparameter choices had the option of relying on squared numbers, log2 numbers, or a total number of features to focus exclusively on the truly important features. This ensures that the model captures only the most important features and incorporates only the most important in the resulting feature importance scores to derive business impacts. The model selected and its most important features are illustrated in Figure 14.

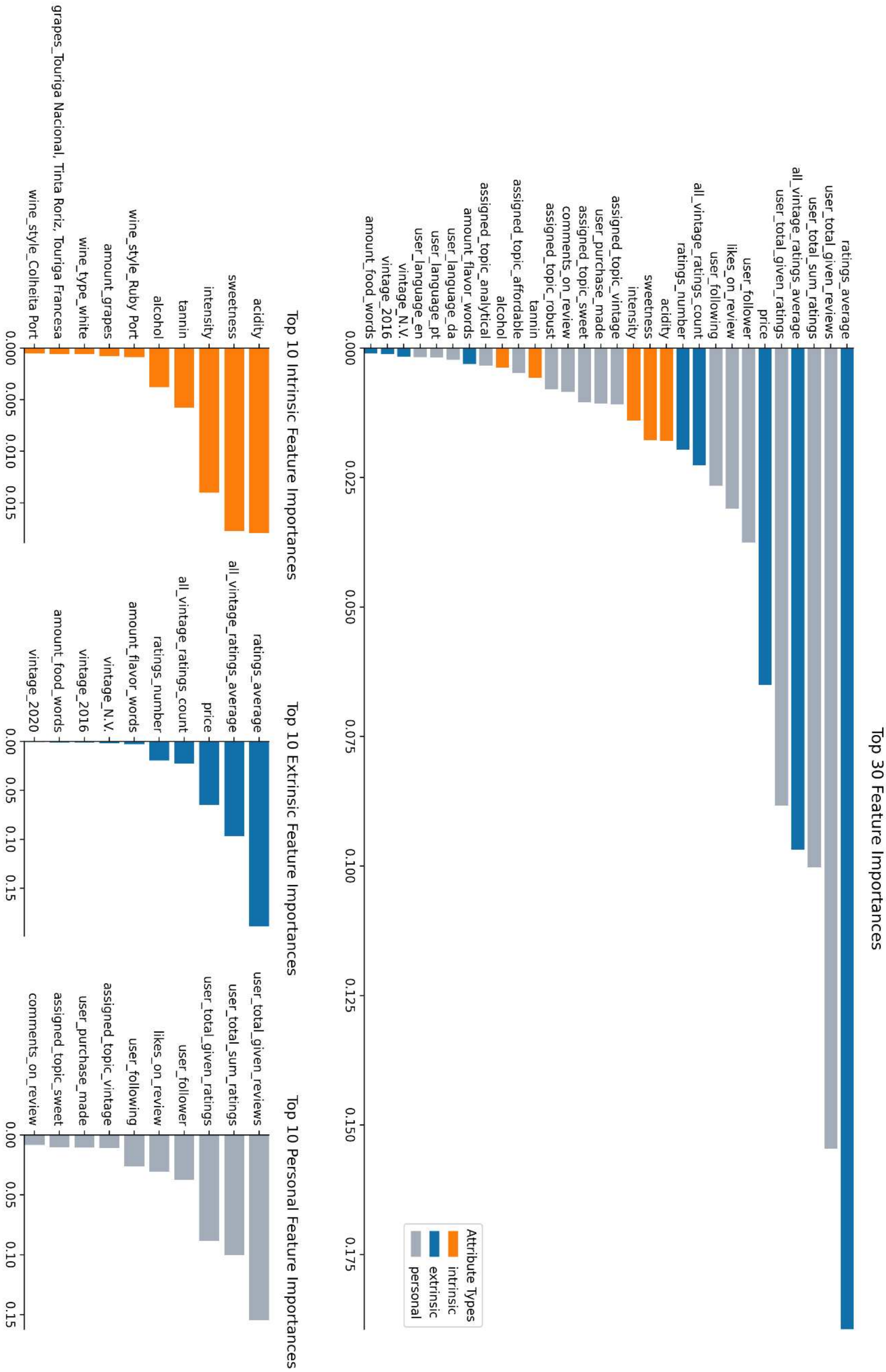


Figure 14: Gradient Boosting Feature Importances by Product Attribute Category

The analysis of this graph reiterates the significant impact of peer ratings, which are supported by ratings averages and all vintage ratings averages. Encouraging consumer participation in providing feedback is essential to determine overall customer satisfaction. Additionally, high values of 'user_total_given_reviews', 'user_total_given_ratings', and 'user_total_sum_ratings' demonstrate the significance of user engagement, further affirming previous findings. The increasing importance of user followers underscores the need to actively engage users to contribute and rate.

While focusing on intrinsic product factors, the effect on accurately determining customer ratings is relatively minor. However, sensory perception remains highly significant in this category, except for the reduced importance of intensity. This emphasizes once again the greater importance of these product characteristics in comparison to the others within this group, such as 'having_wine_style_description'. Nevertheless, it is worth noting that the performance of feature importance highlights a larger number of vintage years, indicating that this factor holds relevance beyond just Port Wine's dominance. The significant importance placed on the assigned topic of vintage reinforces the idea that the production year has a stronger influence compared to other inherent product characteristics.

Finally, the distinct group analysis corroborates the significance of price as an external product attribute. While the subject of affordability remains unproven, the importance of price strongly indicates potential policy decisions. In addition, it is worth noting the visual aspect of the assigned topic for vintages. It could provide an opportunity to identify trends and promote the vintage of wines, as proven by the emerging importance of some vintages in the extrinsic features. Further analysis to derive a trend would be essential.

4.2.4 XGBoost

Having explored the nuances of Gradient Boosting and its impact on feature importance, attention now shifts to XGBoost. Similar to the previous models, XGBoost was initially fitted with default parameters, yielding a training accuracy of 61% and a testing accuracy of 47%. The ROC-AUC showed a revealing effect with a difference between the training value of 92% and the test value of 74%. These discrepancies suggest overfitting and a lack of ability to generalize the data. Adjustments were made to the hyperparameter, resulting in a more balanced training accuracy of 53% and testing accuracy of 47%, paving the way for a robust model capable of drawing reliable business conclusions from its feature importance analysis. The ROC AUC approached 76% for training and 75% for testing.

To reach this performance, the number of estimators was set to a range that wasn't too low to enable the model to detect complex patterns in consumer rating predictions. Having more weak learners impacts the prediction and makes it more aware of the complexity. However, it wasn't set too high to prevent overfitting and the loss of the ability to generalize the prediction outcomes for the consumer ratings. This ability was also aided by a gradual learning rate. As a result, this approach enabled effective handling of imbalanced classes, as the lower learning rate moderated the weighting for misclassification. Additionally, the model captures the presence of this class over time. This resulted in a more resilient model that was further reinforced by subsampling, which continually introduced new data points through bootstrapping. Consequently, the model is prevented from becoming too familiar with the data, thereby reducing variance and resulting in more generalizable outcomes.

Given the dataset's highly dimensional nature of features, a high value for lasso regularization was utilized to select only the most crucial features while penalizing overly complex models. Consequently, only the most important features were included in the model. This also increased the interpretability, as the model only relied on the most important product features. Lastly, the decision was made to adopt a conservative approach when designing the tree model to prevent overfitting and maintain simplicity while still capturing complex patterns. The hyperparameter tuning also leveraged XGBoost's ability to use GPUs, thus reducing processing time. The most important features of this classification task are plotted in Figure 15.

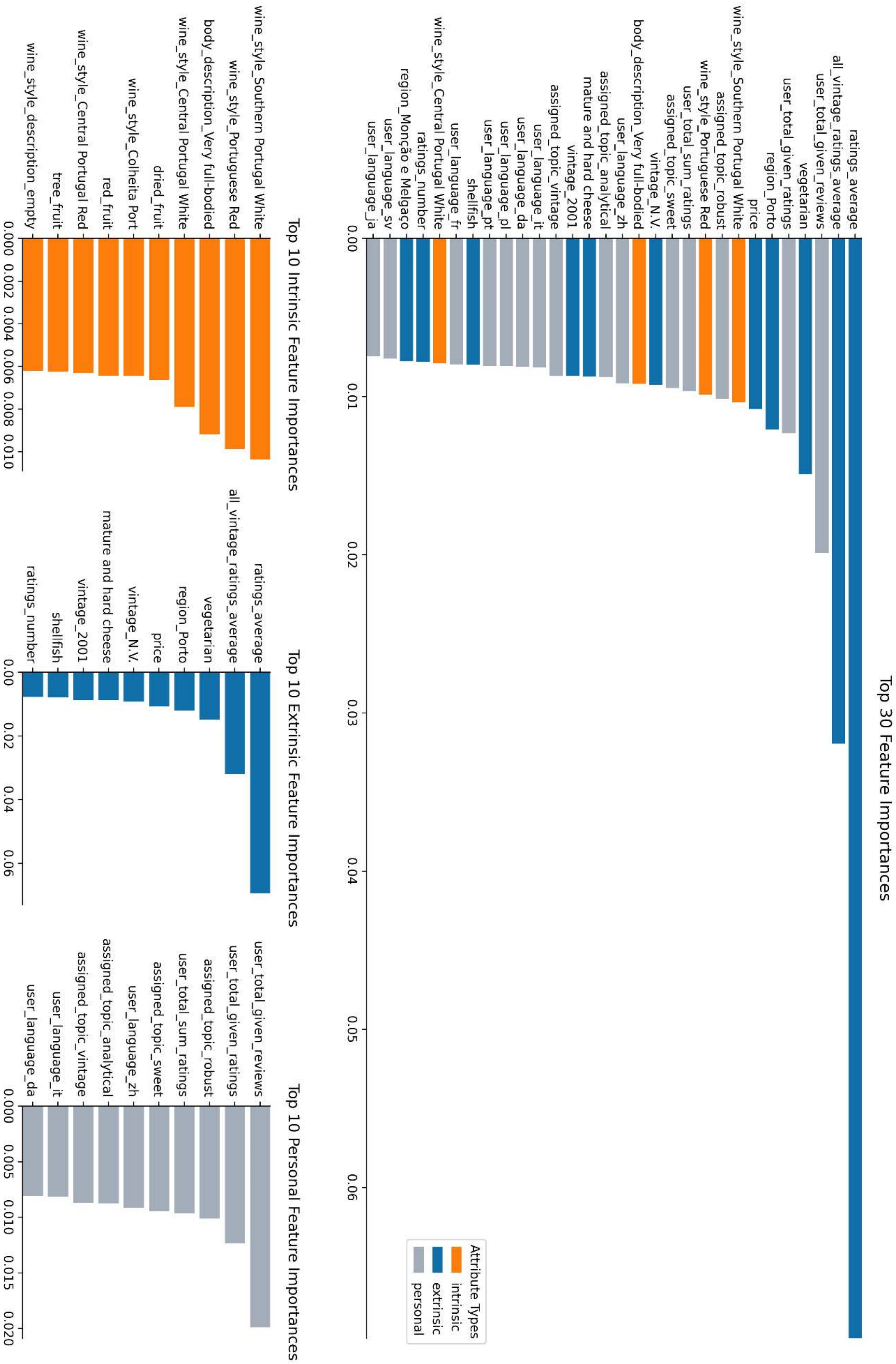


Figure 15: XGBoost Feature Importances by Product Attribute Category

According to Figure 15, the average importance of unique features in this model is significantly lower compared to the other models. The high importance score of 'ratings_average' and 'all_vintage_ratings_average' highlights the critical role of consumer participation. Furthermore, 'user_total_given_reviews' emerges as a key factor in accurately classifying consumer ratings, emphasizing the value of encouraging user reviews. An important observation is the growing significance of the topics mentioned in the comments, such as 'assigned_topic_robust' and 'assigned_topic_sweet'. This suggests a possible shift in focus towards the qualitative aspects of comments. Encouraging more detailed and topic-specific reviews could be advantageous.

Moreover, the prominence of features like 'Southern Portugal whites', 'Portuguese reds', and 'Portugal whites' in the intrinsic product category reinforces the Portuguese prestige. This is particularly noteworthy as the dataset is dominated by Porto and Madeira wines. Furthermore, the significance of fruit characteristics is apparent, as this type of description is present in three different variations within the top 10 most important features. The frequency of this occurrence may suggest a tendency to generalize about fruits to create more universal descriptions by accumulating data. This abundance could be used to carry out a specific analysis for wines with fruity notes.

Regarding extrinsic wine attributes, it is of interest to note the increasing importance of food recommendations. The high importance of the broad category 'vegetarian' may suggest a need for simplification to avoid information overload. While price remains a significant factor, its relative importance seems reduced, as indicated by the absence of the 'affordable' topic among the top personal features. There's also an increased focus on detail-oriented aspects like 'robust' or 'vintage'.

In terms of personal attributes, the language used by the user has become increasingly crucial, revealed by the presence of Chinese, Danish, and Italian as important features. It is interesting to note that although all comments are written in English, the nationality of the commenters seems to influence peer evaluation. Additionally, comments related to sensory perception (such as robust or sweet) are given more weight than those related to price, indicating a high level of appreciation for sensory experiences.

4.3 Business Insights

Having created four robust machine learning models to perform a predictive classification task on the consumer review ratings, revealed a common set of important features. These pivotal factors include consumer participation, user engagement in comments and reviews, sensory evaluation of wines, pricing strategies, and the ability to generalize effectively. Consequently, these insights are instrumental in addressing a critical trade-off: on one hand, there is a need to furnish ample information to cater to the preferences and uncertainties of risk-averse wine consumers, particularly highlighting the unique attributes of Portuguese wines. On the other hand, it is equally crucial to streamline this information to prevent overwhelming customers with excessive details.

To increase consumer participation and encourage a high volume of quality ratings, incentives need to be developed. One method is to introduce a leaderboard that ranks users based on their ratings, fostering a competitive spirit among them. In addition, the rating process can be simplified by providing specific star rating suggestions in words (e.g., "good taste" equals 4.0 stars). Furthermore, implementing review campaigns can remind consumers who have made a purchase or viewed certain wines to leave a rating. Another possibility could be to limit the number of views for wines until a certain number of ratings have been made, or if the user is not registered. By implementing these recommendations, the decision-making process for consumers will be streamlined, focusing attention on the information most influential in driving consumer rating classifications.

Building on these recommendations for increased participation, the next phase focuses on enhancing user interaction to increase engagement and strengthen the sense of community value. This can be done by introducing an award system that classifies users according to their engagement, which includes comments, reviews, ratings, likes, and followers. Possible ranks for reviewers could include Tasting Novice, Enthusiast Explorer, Sommelier Apprentice, or Wine Master. These ranks can be displayed next to the user's name when they make a review. This competitive spirit encourages consumers to create more reviews. Furthermore, this increases the credibility of reviews posted by of reviews made by expert levels. A possible representation is depicted in Figure 16.

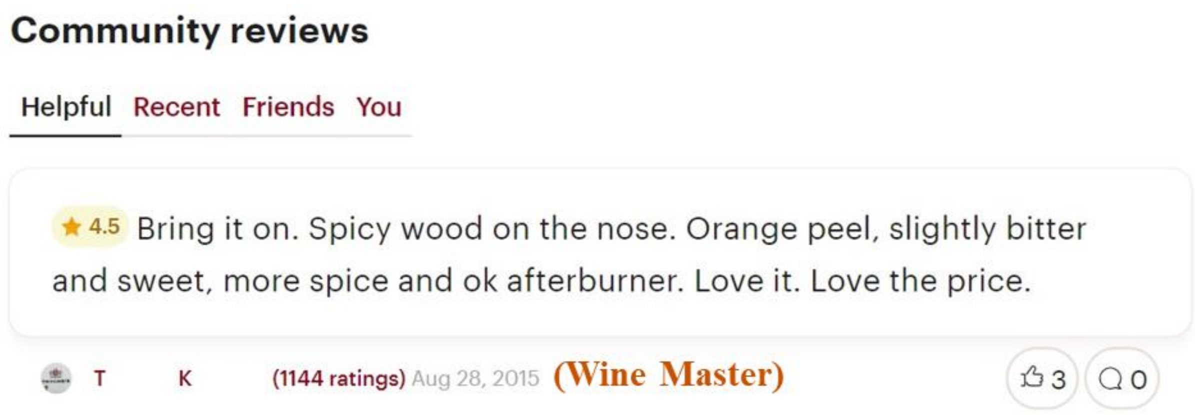


Figure 16: Illustration of Reviewer Ranking, graphic from: (Vivino, 2023b) (own annotation)

This approach incentivizes users by providing digital recognition and also enhances their credibility in assessing information for other users. By incorporating this into the previously introduced leaderboard, a loyalty program could be established to reward frequent user engagement. The distinctiveness of Portuguese wine could be utilized by offering supplementary materials for loyal users, such as exclusive videos showcasing specific regions. Moreover, simplifying the process of writing wine reviews could be achieved by supplying pre-written templates crafted by experts in the industry. Lastly, merging ratings and reviews could create an intangible wine cellar comprising purchased, rated, and reviewed wines, allowing for personalized recommendations for wines based on individual preferences and gathering relevant data on active users. These approaches consolidate existing user information more concisely and credibly while catering to their interest in Portuguese wine.

With these approaches set to enhance user engagement, the subsequent phase introduces a new pricing description strategy. This strategy could include a key indicator comparing the price with the average rating. This KPI would serve as a tool to inform users about the cost-effectiveness of different wines for more informed decision-making. Such a regulator describing the relationship between price and rating is shown in the following figure

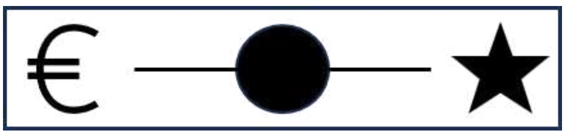


Figure 17: Possible scale for comparing the price with the average rating

To further simplify the pricing assessment for consumers, introducing a descriptive pricing classification could be beneficial. This classification could be linked to a value-added pricing strategy that highlights the unique attributes and rich history of Portuguese wines, distinguishing them from conventional options. For example, emphasizing the heritage, grape varieties, or production methods unique to Portuguese wines could justify their pricing and enhance their perceived value.

Finally, implementing a ‘wine watchlist’ feature would allow users to track price changes. Users could receive notifications when the price of a wine on their watchlist drops, adding a dynamic aspect to the purchasing process. This feature not only keeps consumers engaged but also provides them with timely information to take advantage of price fluctuations. This comprehensive pricing strategy aims to streamline the process of assessing wine prices while offering detailed and specific information about Portuguese wines. It strikes a balance between providing essential pricing information and enhancing the overall consumer experience with additional context and value-based insights.

Alongside refining pricing strategies, the next crucial step involves prioritizing the sensory experience in wine descriptions, as illustrated in Table 2. This approach involves highlighting aspects such as taste and sweetness rather than relying on descriptors like grapes, wine style, or vintage. By integrating sensory terms into the general description alongside the average rating or vintage information, the overall significance and user-friendliness of the description can be enhanced. To ensure effective communication and comparability across various wines and regions, standardizing this methodology is necessary. This can be achieved using familiar comparisons and quantifiable terms that are commonly understood by a broad audience (e.g., sweet matched directly with fruity). This improved representation and simplification can help to focus more on the informative parts of the intrinsic wine product attributes.

Having focused on enhancing sensory descriptions, the final step recommends a broader approach to data generalization, as these tend to be more impactful in feature selection compared to highly specific ones. For example, specific regional information, though important, is often less impactful due to its detailed nature. To address this, leveraging broader classifications like the Denomination of Controlled Origin (DOC) and Regulated Designation of Origin (PDO) can be effective. Utilizing these classifications not only simplifies the information but also capitalizes on their prestige. Possible improvements are illustrated in Figure 18.

Figure 18: Possible improvements for a more effective wine description graphic from: (Vivino, 2023b) (own annotation)

5 Conclusion

The given recommendations aim to address the purpose of this study by managing the trade-off between providing sufficient information for informed decision-making and maintaining precision to avoid overwhelming information. To address this question, a predictive classification task was developed using several machine learning algorithms and statistical concepts, namely Random Forest, Gradient Boosting, AdaBoost, XGBoost, and Latent Dirichlet Allocation. The framework for implementing product attributes by categorizing them into intrinsic, external, and personal categories provides further insights into the distribution of various components that impact consumer satisfaction, measured in this case by consumer star ratings.

Overall, chapter 4.3 of the study reveals that several components, namely consumer participation, user interaction, pricing strategy, and generalization, need to be promoted to provide meaningful business insights. All models, with their different specifications as stated in Chapter 2.3 reveal a broad, overall similar picture for the stated focus areas. Nonetheless, it is crucial to note that the recommendations provided may not guarantee success or yield superior information. They ought to be considered as indications to be tested alongside more robust methods such as A/B tests. Furthermore, it is significant to acknowledge that they may counteract each other and ultimately diminish the perception and evaluation of information rather than enhance it.

The literature-derived framework reveals interesting insights, as all models show the same tendency: the impact of feature importance on correctly predicting consumer rating classes is always smaller compared to the other categories across all four models. This broad overall pattern suggests that researchers who perceive their impact as less significant are validated (Carpenter et al., 1994). In terms of wine, it is somewhat surprising that the experience is conveyed through a set of product attributes within this category (Robertson et al., 2018).

Shifting the focus to extrinsic attributes, the primary impact is on the average rating, which determines the specific consumer review rating. It was a conscious decision to incorporate this rating, as it is a crucial feature of the current Vivino platform. Furthermore, users can filter for a specific average rating. Price is the second-most important factor in accurately predicting classes. Overall, this provides evidence of the significance of extrinsic product attributes (Richardson et al., 1994). Their prominence when compared to other categories also demonstrates their ability to offer supplemental information for effective decision-making (Nelson, 1970; Akdeniz et al., 2013).

Incorporating personal characteristics from comments has a positive effect on peer perception, in addition to the ratings (Kim et al., 2014). This category mainly served as support for observations regarding intrinsic and extrinsic product categories. For example, comments provided information on price, consumer participation, and sensory perception. Including additional information was essential as it provided valuable, sometimes hidden insights (Luo et al., 2008).

In summary, categorizing product features into smaller groups has significantly strengthened the predictive analysis, allowing for the derivation of tangible business impacts. Although the findings provide a clear picture of which elements to focus on to convey information, some points require careful consideration and offer opportunities for further research.

One potential limitation is the reliance on feature importance. Using feature importance as an indicator may result in a correlation effect rather than real causality. This could indicate that some product features are simply correlated without any causal inference. There may also be confounding variables influencing the association between the rating classification and the feature. These effects may not be observable and could lead to incorrect recommendations. This problem was addressed in this study by utilizing various robust models, but there remains an unmeasured risk. Therefore, this could be addressed through randomized experiments in future research.

Another limitation arises in the calculation of feature importance. This study relied on the mean decrease in impurity because the purpose was only to get an indication of its importance. However, this approach could introduce bias for product attributes with multiple distinct labels. To mitigate this limitation, further research could be conducted by relying on alternative feature permutation techniques.

Finally, it is important to consider the ROC-AUC used as an evaluation matrix for the models. This metric does not provide information on specific types of errors, such as false positives and false negatives. Its purpose was to determine the models' ability to distinguish rather than their accuracy. Nonetheless, a central limitation is that the accuracy is relatively low for all models, while ROC-AUC is superior to random guessing. The imbalance among the nine classes led to this issue, as previously stated. Additional research is necessary, indicating a necessity for a web-scraping process that acquires low-rated wines for improved data selection. Regrettably, this could not be executed for the Portuguese excellent wine. Many experiments with down-sampling and upsampling were performed in this study, but none showed any improvement since doing so caused specific characteristics to be lost.

6 Bibliography

Aaker, D. A., & Jacobson, R. (1994). The financial information content of perceived quality. *Journal of Marketing Research*, *31*(2), 191. https://doi.org/10.2307/3152193

Vivino. (2023a). *About Vivino.* https://www.vivino.com/about

Vivino (2023b). Vivino. Retrieved October 6, 2023, from https://www.vivino.com/api/wines/

Akdeniz, M. B., Calantone, R. J., & Voorhees, C. M. (2013). Signaling quality: An examination of the effects of Marketing- and Nonmarketing-Controlled Signals on perceptions of automotive brand quality. *Journal of Product Innovation Management*, *31*(4), 728–743. https://doi.org/10.1111/jpim.12120

Andaleeb, S. S., & Conway, C. (2006). Customer satisfaction in the restaurant industry: an examination of the transaction-specific model. *Journal of Services Marketing*, *20*(1), 3–11. https://doi.org/10.1108/08876040610646536

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y. J., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley, P. (2020). Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657. https://doi.org/10.1145/3351095.3375624

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent dirichiet allocation. In *The MIT Press eBooks* (pp. 601–608). https://doi.org/10.7551/mitpress/1120.003.0082

Boehmke, B., & Greenwell, B. (2019). *Hands-On Machine Learning with R.* In *Chapman and Hall/CRC eBooks*. https://doi.org/10.1201/9780367816377

Bradford, S. (1983). *The story of Port: The Englishman's Wine*. Christie's Wine Publications.

Brechan, I. (2006). The different effect of primary and secondary product attributes on customer satisfaction. *Journal of Economic Psychology*, *27*(3), 441–458. https://doi.org/10.1016/j.joep.2005.10.003

Breiman, L., & Ihaka R. (1984). *Nonlinear Discriminant Analysis Via Scaling and Ace* (Report No. 40). Yale University. Retrieved from https://digitalassets.lib.berkeley.edu/sdtr/ucb/text/40.pdf

Breiman, L. (1996a). *Out-of-bag estimation*. University of California. Retrieved from https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf

Breiman, L. (1996b). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Breiman, L. (1996c). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, *24*(6). https://doi.org/10.1214/aos/1032181158

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3). https://doi.org/10.1214/ss/1009213726

Campbell, C., Ray, G., & Muhanna, W. A. (2005). Search and collusion in electronic markets. *Management Science*, *51*(3), 497–507. https://doi.org/10.1287/mnsc.1040.0327

Carpenter, G. S., Glazer, R., & Nakamoto, K. (1994). Meaningful Brands from Meaningless Differentiation: The Dependence on Irrelevant Attributes. *Journal of Marketing Research*, *31*(3), 339–350. https://doi.org/10.1177/002224379403100302

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Choudhury, V., Hartzel, K. S., & Konsynski, B. R. (1998). Uses and Consequences of Electronic Markets: An Empirical investigation in the aircraft parts industry. *Management Information Systems Quarterly*, *22*(4), 471. https://doi.org/10.2307/249552

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(sici)1097-4571(199009)41:6

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Lecture Notes in Computer Science* (pp. 1–15). https://doi.org/10.1007/3-540-45014-9_1

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1702.08608

Drucker, H., Schapire, R. E., & Simard, P. (1993). Boosting performance in neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, *07*(04), 705–719. https://doi.org/10.1142/s0218001493000352

Efron, B., & Tibshirani, R. (1994). An introduction to the Bootstrap. In *Chapman and Hall/CRC eBooks*. https://doi.org/10.1201/9780429246593

Enneking, U., Neumann, C., & Henneberg, S. (2007). How important intrinsic and extrinsic product attributes affect purchase decision. *Food Quality and Preference*, *18*(1), 133–138. https://doi.org/10.1016/j.foodqual.2005.09.008

Everson, R., & Fieldsend, J. E. (2006). Multi-class ROC analysis from a multi-objective optimisation perspective. *Pattern Recognition Letters*, *27*(8), 918–927. https://doi.org/10.1016/j.patrec.2005.10.016

Faria, S., Lourenço-Gomes, L., Gouveia, S., & Rebelo, J. (2020). Economic performance of the Portuguese wine industry: a microeconometric analysis. *Journal of Wine Research*, *31*(4), 283–300. https://doi.org/10.1080/09571264.2020.1855578

Freund, Y., Schapire, R., & Abe, N. (1990). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence,* 14(771-780), 1612. Retrieved from http://www.yorku.ca/gisweb/eats4400/boost.pdf

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/s0167-9473(01)00065-2

Géron, A. (2023). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.* O'Reilly Media.

Gokcekus, O. & Nottebaum, D (2011). The Buyer S Dilemma: Whose Rating Should a Wine Drinker Pay Attention To?. *American Association of Wine Economists*. Retrieved from https://wine-economics.org/wp-content/uploads/2012/10/AAWE_WP91.pdf

González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, *64*, 205–237. https://doi.org/10.1016/j.inffus.2020.07.007

Grover, V., Lim, J., & Ayyagari, R. (2006). The dark side of information and market efficiency in E-Markets. *Decision Sciences*, *37*(3), 297–324. https://doi.org/10.1111/j.1540-5414.2006.00129.x

Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, *2*(3), 349–360. https://doi.org/10.4310/sii.2009.v2.n3.a8

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning. In *Springer series in statistics*. https://doi.org/10.1007/978-0-387-84858-7

Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition (Vol. 1, pp. 278-282).* https://doi.org/10.1109/ICDAR.1995.598994

Horowitz, I., & Lockshin, L. (2002). What Price Quality? An Investigation into the Prediction of Wine-quality Ratings. *Journal of Wine Research*, *13*(1), 7–22. https://doi.org/10.1080/0957126022000004020

Instituto da Vinha e do Vinho. (2023, November 15). *Vinhos e Aguardentes de Portugal*. Lisbon: Instituto da Vinha e do Vinho.

International Organisation of Vine and Wine. (2023a, November 15). *Data discovery report.* Retrieved from https://www.oiv.int/what-we-do/data-discovery-report?oiv

International Organisation of Vine and Wine. (2023b). *World Wine Production Outlook*.

Johnson, R., & Zhang, T. (2014). Learning nonlinear functions using regularized greedy forest. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(5), 942–954. https://doi.org/10.1109/tpami.2013.159

Kim, H., Park, Y. H., Bradlow, E. T., & Ding, M. (2014). PIE: A Holistic Preference Concept and Measurement model. *Journal of Marketing Research*, *51*(3), 335–351. https://doi.org/10.1509/jmr.10.0406

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 530-539). Association for Computational Linguistics. https://doi.org/10.3115/v1/E14-1056

Laudato, G., et al. (2020). Identification of R-peak occurrences in compressed ECG signals. In *2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA)* (pp. 1-6). IEEE. https://doi.org/10.1109/MeMeA49120.2020.9137207

Levitt, T. (1983). The globalization of markets. *The International Executive, 25*(3), 17-19. https://doi.org/10.1002/tie.5060250311

Luo, L., Kannan, P., & Ratchford, B. T. (2008). Incorporating subjective characteristics in product design and evaluations. *Journal of Marketing Research*, *45*(2), 182–194. https://doi.org/10.1509/jmkr.45.2.182

Marr, B. (2021, May 7). Vivino: Choose your next great wine With big data and artificial intelligence. *Forbes*. https://www.forbes.com/sites/bernardmarr/2021/05/07/vivino-choose-your-next-great-wine-with-big-data-and-artificial-intelligence

Mayson, R. (1998). *Portugal's wines & wine-makers: Port, Madeira & Regional Wines*. Wine Appreciation Guild, The Limited

Menghini, S. (2015). The new market challenges and the strategies of the wine companies. *Wine Economics and Policy*, *4*(2), 75–77. https://doi.org/10.1016/j.wep.2015.11.003

Mimno, D., & McCallum, A. (2007). Organizing the OCA: Learning faceted subjects from a library of digital books. *In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '07)* (pp. 376-385). Association for Computing Machinery. https://doi.org/10.1145/1255175.1255249

Muthukrishnan, A. V., & Kardes, F. R. (2001). Persistent preferences for product attributes: the effects of the initial choice context and uninformative experience. *Journal of Consumer Research*, *28*(1), 89–104. https://doi.org/10.1086/321949

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*. https://doi.org/10.3389/fnbot.2013.00021

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, *78*(2), 311–329. https://doi.org/10.1086/259630

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June). Automatic evaluation of topic coherence. In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics (pp. 100-108). Retrieved from https://aclanthology.org/N10-1012

Oliver, R. L. (1977). Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of Applied Psychology*, *62*(4), 480–486. https://doi.org/10.1037/0021-9010.62.4.480

Optuna. (2023). Optuna: A hyperparameter optimization framework. Retrieved from https://optuna.readthedocs.io/en/stable/

Phillips, R. (2002). *A short history of wine*. Harper Perennial.

Rao, A. R., & Monroe, K. B. (1989). The effect of price, brand name, and store name on buyers' perceptions of product quality: an Integrative review. *Journal of Marketing Research*, *26*(3), 351. https://doi.org/10.2307/3172907

Richardson, P., Dick, A., & Jain, A. K. (1994). Extrinsic and intrinsic cue effects on perceptions of store brand quality. *Journal of Marketing*, *58*(4), 28. https://doi.org/10.2307/1251914

Ridgeway, G. (1999). The state of boosting. *Computing science and statistics*, 172-181. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1aac6453fbb8333ee638b6d8b2bb2aff06c3654b

Robertson, J., Ferreira, C., & Botha, E. (2018). The influence of product knowledge on the relative importance of extrinsic product attributes of wine. *Journal of Wine Research*, *29*(3), 159–176. https://doi.org/10.1080/09571264.2018.1505605

Robinson, J. (2006). The Oxford companion to wine. In *Oxford University Press eBooks*. https://doi.org/10.1093/acref/9780198609902.001.0001

Saar-Tsechansky, M., & Provost, F. (2007). Handling Missing Values when Applying Classification Models. *Journal of Machine Learning Research*, *8*(57), 1623–1657. https://pages.stern.nyu.edu/%7Efprovost/Papers/missing.pdf

Salton G. J., & McGill M. (1983). *Introduction to modern information retrieval*. CiNii Books. http://ci.nii.ac.jp/ncid/BB03535344

scikit-learn developers. (2023). *scikit-learn: Machine learning in Python*. Retrieved from [https://scikit-learn.org/stable/]

Simonson, I. (2015). Mission (Largely) accomplished: What's next for consumer BDT-JDM researchers? *Journal of Marketing Behavior*, *1*(1), 9–35. https://doi.org/10.1561/107.00000001

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. In *Handbook of Statistics* (pp. 303–329). https://doi.org/10.1016/s0169-7161(04)24011-1

Symmank, C. (2018). Extrinsic and intrinsic food product attributes in consumer and sensory research: literature review and quantification of the findings. *Management Review Quarterly*, *69*(1), 39–74. https://doi.org/10.1007/s11301-018-0146-6

Wang, S., Zheng, S., Xu, L., De-Zheng, L., & Meng, H. (2008). A literature review of electronic marketplace research: Themes, theories and an integrative framework. *Information Systems Frontiers*, *10*(5), 555–571. https://doi.org/10.1007/s10796-008-9115-2

Westbrook, R. A. (1980). A rating scale for measuring product/ service satisfaction. *Journal of Marketing*, *44*(4), 68–72. https://doi.org/10.1177/002224298004400410

Zelaya, C. V. G. (2019). Towards Explaining the Effects of Data Preprocessing on Machine Learning. *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. https://doi.org/10.1109/icde.2019.00245

Zhang, T., & Yu, B. (2005). Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, *33*(4). https://doi.org/10.1214/009053605000000255