Consulting Report for Wine on Water

**Improving White Wine Quality**

**&**

**Inventory Selection through Advanced Analytics**

**Prepared by Group 2-C**

**Members:  
Morgan Drouillard  
Sathvika Kokku  
Joakim Kristensen  
Lakshmi Naga Sindhura Adipati**



**Table of Contents**

[Executive Summary 3](#_Toc165308851)

[Introduction 4](#_Toc165308852)

[Client Organization Background 4](#_Toc165308853)

[Current Analytical Maturity Assessment 5](#_Toc165308854)

[DELTA Model 5](#_Toc165308855)

[Current Business Process Issues 7](#_Toc165308856)

[Business Analytics Solution 8](#_Toc165308857)

[Overview of the Dataset 8](#_Toc165308858)

[Data Exploration 10](#_Toc165308859)

[CHAID Model 14](#_Toc165308860)

[Implementation 16](#_Toc165308861)

[Implications 17](#_Toc165308862)

[Conclusion 18](#_Toc165308863)

[References 19](#_Toc165308864)

## **Executive Summary**

Wine on Water is a boutique wine retailer and a valued member of Cru Hospitality Group in Tampa. They want to improve their position in this competitive wine market by using business analytics to choose the best wines for their customers and make them happy. However, they have limited space to store all the different types of wines they want to sell. To solve this problem, they want to improve their inventory selection process, grow their business and increase their customers' satisfaction. So, in this project our team worked with Wine on Water to help them improve their business by selecting some of the best white wines based on their chemical attributes.

For our analysis we mainly focused on two data mining techniques which we found are the most suitable when it comes to identifying which chemical attributes would make a high-quality wine. Finding a high-quality wine based on these data driven approaches not only contributes to higher sales for the company but it also helps the company in managing the inventory by purchasing only the wines which have high quality which in return increases the customer satisfaction.

Throughout the project, we have encountered several challenges that needed a lot of attention and problem-solving skills. One of our key challenges was identifying and selecting the most suitable data mining models that understands the complexity of data and gives an accurate output to draw meaningful conclusions from it.

Additionally, we aimed to present our project in a way that could be easily understood by everyone, even those without technical expertise. However, interpreting most of our analysis in layman's terms with easy-to-understand language was a bit of a challenge. Nevertheless, we were able to tackle it through good team efforts and by working in harmony.

Our analysis mainly focuses on using predictive modeling techniques like regression and CHAID model to make predictions related to higher quality wines. Through our regression analysis and CHAID model we were able to determine the chemical attributes that had heavily impacted the quality of the wine and we were also able to identify the best combination of these chemical attributes and their proportion to make a high-quality wine. When it comes to implementation, we were able to make an analysis which can be easily implemented using proper data mining tools and techniques mentioned in the report. At the end we recommend Wine on Water to take things step by step, making sure all their data is organized and their staff knows how to use the new tools to appropriately implement this data driven approach in the company.

## **Introduction**

In today's fiercely competitive retail landscape, making data-driven decisions has become imperative for businesses to thrive. We observed that while information and analysis about red wines can be found easily on the web, it is not as easy to find the same for white wines, which sets us apart from other projects. Our project, Wine on Water, aims to use advanced analytics to enhance the quality of white wine and optimize inventory selection based on the quality of the wines. This report provides an in-depth analysis of the organization's current analytical maturity, identifies critical business challenges, and presents a tailored analytics solution focused on improving white wine quality and inventory management processes.

## **Client Organization Background**

Wine on Water is a popular place for people in Tampa to buy wine. They have over 150 selections of different wines and spirits, and they focus on supporting wineries that make wine in a sustainable way. Most of their customers come from Tampa and nearby areas. Though it is small in size, Wine on Water plays a significant role in the local wine community, contributing to the cultural blend of the region. They have about 10 people working there, and everyone fulfills different roles, like selling wine, marketing, and keeping track of what they have in stock (inventory management).

In addition to wine, the shop features a curated selection of top distillery and brewery brands from around the world, alongside small batch offerings that add a touch of exclusivity to the collection. The shop's inviting atmosphere includes a rotating by-the-glass list, allowing patrons to enjoy wines on the outdoor patio. Wine on Water also offers its beautiful space for private events and wine tastings, providing an ideal setting for social gatherings and special occasions. Complementing its retail offerings, Wine on Water hosts an extensive series of educational classes and tastings throughout the year. These events provide customers with opportunities to expand their knowledge and appreciation of wines and spirits, enhancing their overall shopping experience. With its diverse selection, inviting ambiance, and commitment to customer education, Wine on Water has become a valued destination for wine and spirits enthusiasts in the local community.

## **Current Analytical Maturity Assessment**

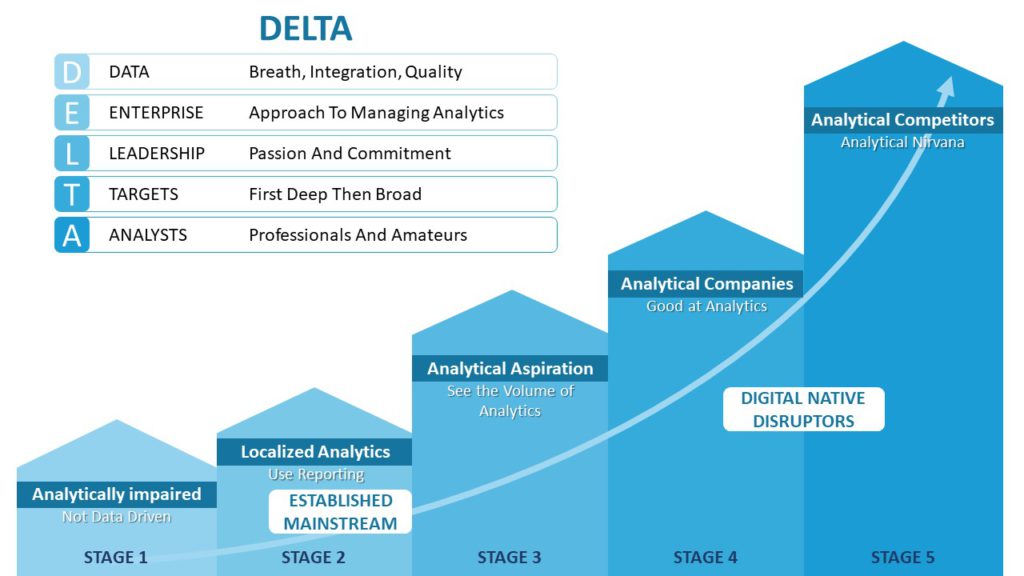
The DELTA Model is a framework developed to assess an organization’s maturity in its analytics and data-driven decision-making capabilities. DELTA is an acronym that stands for Data, Enterprise, Leadership, Targets, and Analysts. Data quality and ease of access are fundamental. It's important to have correct, complete, and easily reachable data, supported by strong management systems. The Enterprise part of the model emphasizes that analytics should be part of all business areas and embedded into main activities. Leadership is key because strong leaders need to support and guide the use of analytics, giving clear directions and resources. Targets are about setting clear and specific goals that analytics can help achieve, making sure these goals fit with the bigger plan of the company. The Analysts part points out the need for skilled people who can understand data, find valuable insights, and communicate these insights well to help make better decisions.

Each of these components represents a crucial aspect that organizations need to address to build a robust analytics program. Here’s a brief overview of each component with respect to our company Wine on Water:

### **DELTA Model**

* **Data**: Wine on Water relies on a Point of Sale (POS) system for most reporting needs, storing data across various systems such as Excel spreadsheets and an email marketing tool. However, there is no dedicated data warehouse, leading to challenges in data integration and quality assurance. As the organization focuses on white wine quality assessment, establishing a centralized data repository for white wine-related data would be beneficial to support advanced analytics initiatives.
* **Enterprise**: With a small team of 10 employees and a limited budget, Wine on Water faces challenges in adopting sophisticated analytics tools. However, there is an interest in exploring customizable solutions to enhance data-driven decision-making processes. A strategic commitment to analytics at the organizational level is crucial for successful implementation and integration of analytics solutions into existing business processes.
* **Leadership**: The absence of a dedicated business analytics team suggests a lack of leadership focus on data-driven decision-making. Leadership buy-in is essential to drive analytics initiatives forward and foster a culture of data-driven decision-making within the organization.
* **Targets**: Clear targets and performance metrics for white wine quality improvement initiatives are lacking. Establishing measurable goals is crucial for tracking progress and evaluating the success of analytics initiatives. Setting targets related to inventory optimization, customer satisfaction, and sales growth would align with the organization's business objectives.
* **Analysts**: Analytics tasks are distributed among staff members who have the requisite knowledge, indicating a decentralized approach. However, building analytical capabilities and fostering a data-driven culture among all employees is vital for long-term success. Providing training and resources to enhance analytical skills would empower employees to leverage data effectively in decision-making processes.

Overall, Wine on Water is at the beginning stages of the business analytics maturity model, primarily operating on descriptive analytics. To advance to higher stages of maturity and realize the full potential of business analytics, the organization should focus on enhancing data integration and quality, fostering leadership buy-in, establishing clear targets, and building analytical capabilities among staff members.



## **Current Business Process Issues**

Wine on Water has some problems that could be fixed with business analytics. Right now, they struggle with things like suboptimal inventory forecasting, a weak grasp on consumer buying behaviors, and ineffective marketing campaigns, all of which directly impact financial performance and customer satisfaction. One of the primary business problems at Wine on Water is inventory management. The business struggles to predict accurately the demand for various wines, often resulting in either excessive or insufficient stock. This misalignment between supply and demand leads to lost sales opportunities and dissatisfied customers who can't find their preferred wines. Additionally, marketing efforts have been somewhat generic, not leveraging customer data to target potential buyers effectively, thus failing to achieve desired engagement or returns on investment.

* **Limited Inventory Optimization**: Despite offering a diverse selection of wines, Wine on Water struggles with inventory optimization due to its small inventory space. Balancing the need for variety with the constraints of physical space poses a challenge in ensuring a comprehensive and appealing selection of wines that align with customer quality expectations.
* **Quality Assessment Reliability**: The organization heavily relies on subjective assessments and past sales data to gauge the quality of wines. However, this approach may not consistently align with broader customer quality expectations. Without objective criteria for assessing wine quality, there is a risk of stocking wines that do not meet customer preferences, leading to potential dissatisfaction and loss of sales.
* **Nature of the Problem**: The primary issues at Wine on Water revolve around inventory management and marketing, with underlying themes in data utilization and customer satisfaction. Addressing these challenges requires a strategic approach that leverages business analytics to optimize inventory selection, improve quality assessment processes, and enhance targeted marketing efforts to better meet customer needs and preferences.

## 

## **Business Analytics Solution**

To address the challenge of limited information and analysis available for white wines, Wine on Water can leverage business analytics methods.

* **Predictive Modeling for Quality Assessment**: Implement regression or classification models to predict white wine quality based on chemical attributes, enabling proactive inventory selection.
* **Inventory Optimization**: Prioritize stocking white wines with chemical profiles aligned with higher quality scores to optimize inventory within space constraints.
* **Customer Satisfaction**: Align inventory with quality predictions to increase customer satisfaction and loyalty, leading to improved sales and profitability.

### **Overview of the Dataset**

The wine quality data set comprises of white variants of wine. This dataset serves as a valuable resource for understanding, analyzing, and predicting the quality of white wine based on a comprehensive set of physicochemical attributes and sensory evaluations. By examining these attributes and evaluations, we can gain insights into the factors influencing wine quality such as acidity levels, sugar content, alcohol concentration, and sensory perceptions like taste and aroma. This information facilitates the development of predictive models and analytical approaches to access and improve the quality of white wine which would help in optimizing production processes and refining quality control measures, ultimately elevating customer satisfaction within the wine industry.

This dataset contains numeric attributes representing various physicochemical attributes of wine such as acidity, sugar content, and alcohol concentration.

* The 'Quality' attribute serves as the target variable, indicating the perceived quality of wine based on sensory evaluations.
* The Quality ratings range from 3 to 9, with higher scores >7 indicating better quality.

|  |
| --- |
| **White Wine Dataset** |
| Number of Instances: 4898 |
| Number of Attributes: 12 |

**Attribute Descriptions**

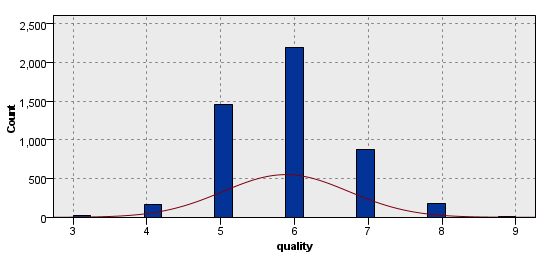
|  |  |
| --- | --- |
| **Attribute** | **Description** |
| Fixed Acidity | The fixed acidity of the wine (g/dm3) |
| Volatile Acidity | The volatile acidity of the wine (g/dm3) |
| Citric Acid | The citric acid content of the wine (g/dm3) |
| Residual Sugar | The residual sugar content of the wine (g/dm3) |
| Chlorides | The chloride concentration of the wine (g/dm3) |
| Free Sulfur Dioxide | The free sulfur dioxide content of the wine (mg/dm3) |
| Total Sulfur Dioxide | The total sulfur dioxide content of the wine (mg/dm3) |
| Density | The density of the wine (g/cm3) |
| pH | The pH value of the wine |
| Sulphates | The sulfate content of the wine (g/dm3) |
| Alcohol | The alcohol content of the wine (% vol.) |
| Quality | The quality rating of the wine, ranging from 3 to 9 (based on sensory evaluations) |
|  |  |



### **Data Exploration**

Before we use our regression model to analyze the data, we must ensure the "Quality" variable scores are normally distributed. Regression data mining techniques rely on the assumption that the dependent variables are normally distributed to make valid statistical inferences. To check for normal distribution, we can use techniques such as visual inspection of histograms, Q-Q plots, or statistical tests like the Shapiro-Wilk test.

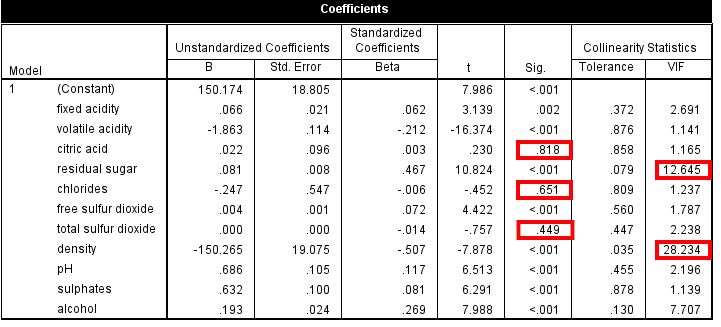
For our data set, a histogram (figure 1) ranging from 3 to 9 was used, along with a distribution curve to verify normality. The highest bars are centered around a score of 6, with a slight left skew, suggesting that this is the most common quality score in the dataset. There's a noticeable decrease in frequency as the scores move away from 6 towards the extremes of 3 and 9.



(Histogram, Figure 1)

The bars for scores 3 and 9 are quite low, which could suggest outliers or a limited range of scores. A data audit was performed, and the outliers were coerced along with nullifying the data extremes.

**Regression Model**



(Entry Regression Model, Figure 2)

To begin the Regression analysis, we needed to examine the possible correlation among the dependent and independent variables. We also want to eliminate variables that do not support the overall model fit. By running an Entry regression model (figure 2) on the unfiltered data, we were able to make the following initial observations:

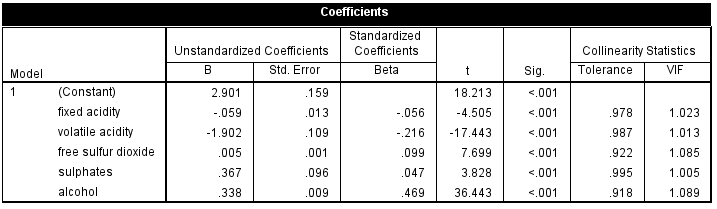
1. Significance (Sig.): Variables like 'volatile acidity', 'residual sugar', 'free sulfur dioxide', 'density', 'pH', 'sulfates', and 'alcohol' are statistically significant predictors of the response variable as indicated by their p-values (<0.001).
2. High Variance Inflation Factor (VIF): 'Density' has a VIF of 28.234, which is much higher than the commonly used threshold of 5 or 10, indicating it may be causing multicollinearity issues. ‘Residual sugar’ has a VIF of 12.645, also putting it about the threshold of 10.
3. Tolerance: 'Density' again shows a tolerance of 0.035, which is lower than the common threshold of 0.1, also indicating a potential multicollinearity problem.
4. Standardized Coefficients (Beta): The standardized coefficients indicate the relative importance of each variable when predicting the outcome. For example, 'alcohol' has a higher positive impact (0.269), and 'volatile acidity' has a strong negative impact (-0.212) on the response variable.

We then decided to filter out the following variables that did not meet the accepted thresholds and re-run the regression analysis to examine if the filtered dataset provided a better model fit. After running a second regression, the “ph.” value showed a Sig score of above 0.1, therefore it was removed and the final model fit was established with the proceeding variables:

* Citric Acid
* Residual Sugar
* Chlorides
* Total Sulfur Dioxide
* Density
* pH

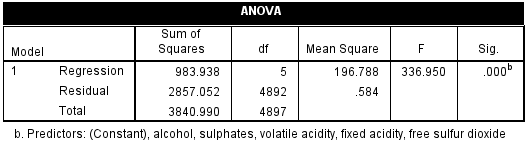
**Regression Analysis**

After running the regression analysis against the filtered data (figure 3)



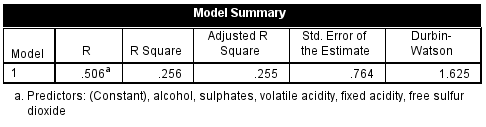
(Filtered Regression Analysis, Figure 3)

The remaining variables are all accepted against the determined thresholds, and we can now move forward with the Regression analysis. We analyzed the ANOVA chart, along with the Model Summary, to determine if the model was statistically significant, and if a model formula could be generated from the coefficients.



(Anuva Chart, Figure 4)

The main value of note is the Significance (Sig) score, otherwise known as the p-value. The p-value associated with the F-statistic is .000b, indicating that the likelihood of observing such a large F-statistic under the null hypothesis (that the model with no predictors fits the data as well as your model) is less than 0.1%. This very low p-value suggests that our regression model provides a significantly better fit to the data than a model with no predictors.



(Model Summary, Figure 5)

After verifying significance, we analyzed the model summary to give us a better understanding of the fit of the regression model.

1. R: The multiple correlation coefficient, here 0.506, indicates a moderate correlation.
2. R Square: This is the coefficient of determination, which represents the proportion of variance in the dependent variable that can be explained by the predictors. Here, it's 0.256, meaning that approximately 25.6% of the variance in the dependent variable is accounted for by the model.
3. Adjusted R Square: This is the R Square adjusted for the number of predictors in the model, providing a more accurate estimate when comparing models with different numbers of predictors. With a value of 0.255, it's very close to the R Square, indicating that the number of predictors is appropriate for the sample size.
4. Std. Error of the Estimate: Also known as the standard error of the regression, this value, 0.764, indicates the average distance that the observed values fall from the regression line. In other words, it's the standard deviation of the residuals or prediction errors.
5. Durbin-Watson: The Durbin-Watson statistic tests the residuals for autocorrelation. Its value ranges from 0 to 4, with values around 2 indicating no autocorrelation. At 1.625, there might be a mild indication of positive autocorrelation in the residuals, although it's not severe. Values significantly less than 2 suggest positive autocorrelation, and values significantly greater than 2 suggest negative autocorrelation.

Based on the metrics, the model is moderately fitting the data with some room for improvement, as indicated by the R-Square and Adjusted R Square values. The presence of autocorrelation as suggested by the Durbin-Watson statistic might warrant further investigation, such as looking into data collection methods that could have contributed to this pattern in the residuals. Overall, we can support model fit, significance, and correlation, therefore we are able to construct a regression equation.

*Predicted Value=2.901+(−0.059×Fixed Acidity) +(−1.902×Volatile Acidity) +(0.005×Free Sulfur Dioxide) +(0.367×Sulphates) +(0.338×Alcohol)*

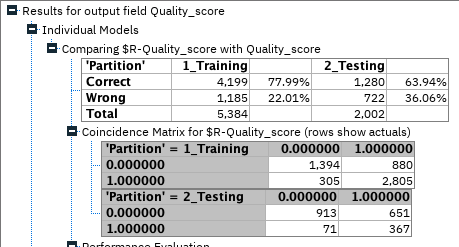
This equation can now be used to make predictions based on the values of the variables included in our model. Each coefficient represents the expected change in the dependent variable for one unit of change in the respective independent variable, holding all other variables constant. The constant term (2.901) represents the expected value of the dependent variable (quality) when all independent variables are set to zero.

### **CHAID Model**

We decided to use the CHAID model for our classification in this dataset. In order to use the model, we chose to filter the variables down to five variables. Our final predictor variables were Volatile Acidity, Chlorides, Total Sulfur Dioxide, Density and Alcohol with quality as our target variable. A 60/40 split partition was added to validate our model’s accuracy. In order to classify the variables that relate to high quality score, we have added an additional quality score column to the original dataset where all the variables that relate to a quality of 7 and above are taken as 1 and all the variables that relate to less than 7 qualities are taken as 0 in the quality score column. To address imbalance in the dataset, we assigned different factors for each of the weights, with 5.0 for Quality Score = 1 (indicating high quality) and 1.0 for Quality Score = 0 (not high quality). This ensured that the data was further balanced for our final analysis.

The outputs from our finalized classification model in the testing partition indicated that there were 76 false negatives and 597 false positives, while it accurately predicted 967 true negative and 362 for true positive. These results highlight (figure 1) our models’ capabilities and limitations in classifying the quality of wine based on the five chosen predictor variables in our dataset.

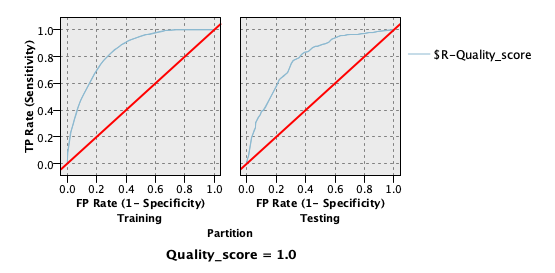
|  |  |  |
| --- | --- | --- |
| CHAID | Accuracy | 63.94% |
| Sensitivity | 83.79% |
| Type-II Error Rate | 16.21% |



(Classification Analysis, figure 6)

We decided to use Receiver Operating Characteristic (ROC) curves to determine the true positive rate against the false positive rate to get a better reading of our dataset. The model's performance was assessed using ROC curves for the training and testing data partitions in SPSS Modeler. The training partition's curve indicates the model's capacity to distinguish between classes, which is shown by being near the upper-left corner of the graph, representing a higher true positive rate and lower false positive rate. The testing partition's curve demonstrates the model's consistency, which remains well above the baseline that indicates a good generalization from training to testing data in our model.

By evaluating our model with ROC curve, we can determine the most optimal classification threshold that meets the models’ operational criteria. This helps maintain a balance between correctly identifying high quality wines and minimizing incorrect classifications based on the variables we have set in the model. Hence, this provides a detailed understanding of the model's performance, combining classification metrics and ROC curve analysis to improve our method for determining wine quality based on our set variables that we used.



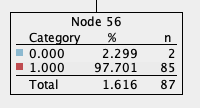
(ROC Curve, figure 6)

Figure 3 presents a CHAID decision tree model visualizing the data segmentation based on sequential variable decisions. Each intersection or node in the diagram signifies a variable from the dataset where data splits occur, moving from one variable to the next as determined by the chi-square test.



(CHAID Model, figure 7)

Figure 4 shows a detailed view of a specific node from the decision tree. This node captures the probability of outcomes with percentage values for each category and a count of observations for each outcome, summed up in the total at the node. This node is the result of a sequence of decisions starting from the top of the tree and moving through the variables quality, alcohol, sulphates, to volatile acidity and has a 97.71% probability to indicate a classification or outcome based on these variables. In other words, this probability suggests a strong likelihood that the conditions of quality, alcohol content, sulphates, and volatile acidity, as represented by this node, lead to a specific classification within the dataset.



(CHAID Model Node 56, figure 8)

## **Implementation**

The implementation process begins with meticulous data collection. Key chemical properties of the wines such as fixed acidity, volatile acidity, free sulfur dioxide, sulphates, and alcohol content form the dataset necessary for predictive analysis. With data in hand, the next step is the application of a regression equation that has been developed to forecast the 'quality' scores of the wines. This equation will be integrated into the existing software system of the wine shop, enabling a seamless and automated prediction process.

Once the 'quality' scores are obtained, these figures will be used to make informed decisions about the wine inventory. Wines that boast higher predicted 'quality' scores will be prioritized, thereby ensuring that the shop's selection reflects wines that are more likely to be appreciated by customers.

Further refining the selection process, the CHAID model serves as a classification tool, distinguishing wines into categories of 'high quality' and 'not high quality.' The practical application of this classification is twofold. Firstly, it assists in inventory categorization, facilitating a strategic approach to stock variety. Secondly, it enhances the customer recommendation system, enabling staff to provide customers with options that are scientifically backed to be of superior quality.

The customer experience is augmented through the development of profiles based on purchase histories and preferences. These profiles, when analyzed in conjunction with the 'quality' classification of wines, allow for tailored recommendations, fostering a personalized shopping experience. Staff play a crucial role in this data-driven strategy. As such, they will undergo training to adeptly use the model predictions to guide customers in their selections. They will also be equipped to educate customers on how the wines are evaluated and recommended based on their chemical attributes, adding a layer of transparency and trust to the customer service experience.

The marketing and sales strategy will pivot around the 'high quality' classification. By spotlighting these wines through promotions and targeted marketing, the shop can attract customers keen on high-caliber wines, thus potentially boosting sales. To close the loop, customer feedback on the wine recommendations will be solicited and analyzed. This feedback is not merely for customer engagement but serves as a critical component in refining and validating the predictive models. The sales data, along with customer feedback, will provide a feedback mechanism to continually enhance the accuracy and effectiveness of the wine selection process.

The adoption of a data-driven strategy in wine selection is not only innovative but positions the shop as a forward-thinking player in the industry. It is a strategy that is grounded in scientific analysis yet aimed squarely at the heart of customer satisfaction. Regular monitoring and updating of the predictive models will ensure that the approach remains dynamic and responsive to the evolving tastes and preferences of the customer base.

### **Implications**

To successfully implement a new business analytics strategy, Wine on Water must consider its significant implications across various dimensions, including managerial, technical, financial, and ethical. From a managerial standpoint, the company must ensure that they have a well-planned and well-communicated strategy that aligns with the broader business objectives. They also need to harmonize their disparate data sources into a centralized repository, which will require a lot of work.

Additionally, the company needs to make strategic decisions on how to allocate their resources within their budget constraints. Finally, they need to get approval from their leaders to endorse the data-centric strategies. On the technical front, the company must ensure that their data remains accurate and of high quality. They need to use pre-processing techniques to ensure data integrity, and they need to choose the right models, such as CHAID classifiers and linear regression models. They also need to have skilled technical personnel to evaluate and deploy these models within existing business processes. Finally, they need to continuously update and retrain their models to maintain relevance.

The financial implications of this strategy are also important. The company needs to evaluate how much they can spend on creating data warehouses, acquiring tools, and training their staff. They also need to weigh these costs against the expected returns and long-term benefits. Finally, ethical considerations are critical. The company needs to ensure that they use customer data responsibly, uphold privacy standards, and use insights ethically to improve customer experience. They must ensure that they do not manipulate customers or use their data inappropriately.

## **Conclusion**

At Wine on Water, the adoption of business analytics is not merely a technological endeavor but a strategic initiative encompassing managerial, technical, financial, and ethical considerations. We began by meticulously preparing our data, we ensured the data was clean and standardized before performing analysis. Our approach involved addressing missing values and standardizing data formats to facilitate accurate model building. Subsequently, we selected tailored models aligned with the shop's business objectives, such as multiple linear regression for predicting wine quality and CHAID classification for categorizing wines based on specific attributes. We rigorously assessed these models, examining metrics like R-squared and accuracy to determine their performance and effectiveness. Once satisfied with the model's performance, we generated outputs that provided valuable insights for decision-making.

Furthermore, we acknowledge the importance of enhancing our analytics capabilities and upholding ethical standards in our operations. As such, we propose a set of recommendations to strengthen our analytical skills and ensure ethical usage of data. We recommend the company to include a phased approach to centralizing data, implementing transparent decision-making processes for resource allocation, and establishing an internal analytics task force led by a dedicated champion to facilitate adoption and cultural change.

Moreover, we recommend them to advocate for partnerships and cost-effective solutions to address technical gaps, alongside the development of financial models to track analytics investments against performance metrics. Furthermore, they must develop a clear data governance policy to uphold ethical standards and regulatory compliance. By constantly incorporating these recommendations into the analytics framework, Wine on Water can stand on the top to unlock the transformative potential of data-driven insights, thereby enhancing the customer experience and fostering sustainable business growth.

## 

## **References**

1. Wine on Water official website:<https://www.wineonwatertpa.com/>

Information regarding the organization's services, offerings, and commitment to providing a diverse selection of wines and spirits was obtained from the official website.

1. UCI dataset on wine quality:https://archive.ics.uci.edu/dataset/186/wine+quality

The dataset provided valuable insights into wine quality attributes, which were utilized for analysis and recommendations regarding inventory selection and customer satisfaction strategies.

-x-