An analysis of underlying relationships between factors related to operating costs and revenue in Australian vineyards.

Abstract

Through a nationwide data set, collected over ten years, we link key variables in determining vineyard operational costs and revenue through the use of XGBoost. We further use a measure of relative importance to show the interrelated nature of these variables and the comparative influence they have on one another. Connections betwen variables is presented through the use of Sankey and Chord diagrams to show the important predictors of revenue and operating costs and their strong interrelatedness. Furthermore, we connect these variables to different wine regions, highlighting the complex influence of location on the use of different resources. With the Australian wine industry being a major contributor to Australia's agricultural sector and economy, this study provides valuable insights into the multifaceted dynamics governing operational costs and revenue, illustrating how factors such as water and fuel use impact operational costs and how different seasonal events affect these operations.

1. Introduction

Strong demands for Australian wine have historically helped to create a thriving industry. However, recent pressures brought on by a loss of tourism and labour due to the COVID-19 pandemic, the global freight crisis, war in Europe, tariffs and rising inflation have negatively affected the industry's outlook (Wine Australia, 2021; Australia, 2021a). The 2021-2022 financial year alone saw a decline of 19% in exports solely due to tariffs (Wine Australia, 2022). A greater understanding of the different underlying conditions leading to improved performance in agricultural productivity and sustainability at scale is key to making data-informed decisions to increase a nation's agricultural sustainability (OECD, 2019). Specifically within the Australian wine and vine industry, there is a need to further understand the driving relationships between resource use and economic output, which can help to determine more cost effective, efficient methods, and to develop benchmarks with local growers (Luke Mancini, 2020).

The potential for new insights into the driving economic forces of the Australian wine industry have manifested in an unprecedented amount of data regarding Australian winegrowing, collected through the Sustainable Winegrowing Australia program. A major part of the insights within this dataset come from the incorporation of operating costs and grape revenue, with environmental and sustainable data. We seek to address both the predictability of operating costs and revenue within the australian winegrowing context and examine their major driving factors to observe linked trends in sustainable practices. As part of this we examine the data to study economic outcomes and their statistical relationships to vineyards' utilisation of resources. We adopt a popular, relatively new machine learning technique, XGBoost, for this analysis because it is able to overcome multicollinearity as well as highlight the level of importance that predictor variables have on response variables (Chen and Guestrin, 2016).

This study is further driven by recent reveiews calling for data-driven studies to show the economic benefits of sustainable practices within the wine industry, specifically winegrowing. While there is evidence to suggest that environmentally sustainable pracitces can reduce costs, increase efficiency, and improve the quality of grapes, more research is needed to numerically demonstrate these benefits across different regions and climates (Baiano, 2021; Mariani and Vastola, 2015; Montalvo-Falcón et al., 2023; Laurent et al., 2021). Futhermore, many different sustainable approaches exist but are often studied in isolation or are limited in their geographical and climatic conditions, restricting their generalisability. We embrace the variation that exists between vineyards and their unique challenges across Australia. Where, vineyard decisions on-the-ground are governed by complex physical forces of a regions' resources, climate, soil and geology, as well as by external pressures such as international market demands, disease and natural disasters (Abad et al., 2021; Cortez et al., 2009; Goodwin I, Jerie P, 1992; Hall et al., 2011; Kasimati et al., 2022; Oliver et al., 2013; Srivastava and Sadistap, 2018).

2. Methods

2.1. Data

Data used in this analysis were obtained from Sustainable Winegrowing Australia (SWA), Australia's national wine industry sustainability program. SWA aims to support grape growers and winemakers in demonstrating and improving their sustainability (SWA, 2022). Data recorded by SWA are entered voluntarily by winegrowers, manually using a web based interface. There are a total of 6049 observations were collected from 2012/2013 to 2021/2022 financial years. Variables recorded by winegrowers are optional. Each vineyard record consists of observations comprising 23 variables reflecting a vineyard's state for the given year (see Table 1). The data was restricted to vineyards that at minimum recorded vineyard size.

Due to the optional and manual recording of data, steps were taken to remove potentially erronous entries. This process first involved discussions with SWA highlighting possible entry errors. At the end of a season any suspect entries, such as a missing fuel-use in a vineyard that recorded the use of tractors, would warrant calling individual vineyards to clarify values and logic within the data. Similarly suspicious entries within the data were first described to viticulturalists for scrutiny before being addressed, either through calling growers for clarification or the removal of an observation due to its unlikely plausability, with most cases suspected of being incorrect units (commonmly litres instead of megalitres of water used) but were not able to be verified.

Due to the nature of XGBoost (eXtreme Gradient Boosting) data was not required to be scaled before used. However some transformations were done, such as multiclass variables being converted to one-hot-encoded variables (the only multiclass variables originally included were year and region). Variables relating to resource consumption, such as water-use were originally divided into whether it was river, dam, or pressurised water but were summed into total water/electricity/diesel/petrol. The source of these variables (such as river, dam, pressurised water) were then converted into binary variables that

Table 1: Summary of variables used in the analysis. The recorded column indicate the number of values that were either greater than zero or that were not missing (see Appendix for more information).

Variable	**Units**	**Number of Classes**	**No. Records**
Water Used	Mega Litres		5846
Diesel	Litres		5585
Biodiesel	Litres		25
LPG	Litres		958
Herbicide Spray	No. Times per year		2026
Year	Class	10	6049
Disease	Class	2	6049
Region	Class	58	6049
Solar	Kilowatt Hours		622
Irrigation Type	Class	20	6049
Petrol	Litres		4309
Slashing	No. Times per year		2290
Yield	Tonnes		5935
Irrigation Energy	Class	16	6049
Area Harvested	Hectares		6049
Electricity	Kilowatt Hours		1014
Insecticide Spray	No. Times per year		1092
Fertiliser	KGs of Nitrogen		795
Fungicide Spray	Times per year		2260
Cover Crop	Class	32	6049
Water Type/Source	Class	39	6049
Grape Revenue	AUD		853
Operating Costs	AUD		853

reflected the presence of a source being used. Other variables that reflected types of operations used such as irrigation-type and cover-crops were also converted to reflect whether a grower simply used these types of systems as opposed to the original format being the specific hectares covered by them. This decision due to a majority of vineyards utilising one source or a second as a backup, with an overwhelming percentage of water/electricity/irrigation prevailing within a single vineyard. The use of a binarisation also meant that importance measures would be better understood as they forced the ensemble to partition by presence or absence of a type of system as opposed to an overly specific number of hectares. This further helped to utilise relative importance for these variables directly to the act of using one system over another. This approach was compare to using the original variables but little difference in model accuracies was found between variables reported as proportion of a type used (i.e the percentage of land covered by drip irrigation), direct units of a type (i.e ML river water used) or as a binary presence/absence. Further details about these variables, their classes and their frequency is available in the Appendix.

2.2. Additional regional data

The variable Region represented one of the 65 Geographical Indicator Regions (GI Region) used to describe unique localised traits of vineyards across Australia (Halliday, 2009; Oliver et al., 2013; SOAR et al., 2008). Each region is explicitly defined under the Wine Australia Corporation Act of 1980 (Attorney-General's Department, 2010). The regional data also expanded to include summary information regarding regions' climate and terrain in the form of minimum, maximum, median and range of elevation. And, temperature and rainfall means alongside extreme heat and cold days; as well as a regions' aridity index. This data was sourced using https://onlinelibrary.wiley.com/doi/full/10.1111/j.1755-0238.2010.00100.x And https://www.wineaustralia.com/growing-making/environment-and-climate/climate-atlas TODO: refs above

2.3. XGBoost

XGBoost is an ensemble method that combines multiple decision trees together to create a more accurate predictive model. The gradient boosting aspect of the ensemble is the use of a loss function to create new decision trees that add to the ensemble, improving its predictive power. The loss function is optimised iteratively to improve upon prior trees (where the loss function can be any convex function), allowing gradient descent to traverse the loss space until no substantive improvements can be made (further detail pertaining to the algorithm is described in the Appendix). Because the loss function is only required to be convex, both classifiers and regressors can be used. Regularisation methods can also be incorporated to help prevent over fitting. This makes XGBoost incredibly versatile and accurate, whilst still being interpretable compared to other machine learning methods (Kisten et al., 2024).

XGBoost analyses were conducted using the XGBoost library (Chen and Guestrin, 2016) in the Python Programming language (G. van Rossum, 1995). It is a method that is widely used within agriculture for yield prediction (D. Mariadass et al., 2022; Li et al., 2024; Ravi and Baranidharan, 2020), but is also highly capable method for financial predictions, even when dealing with multi-domain predictor variables (Zhang et al., 2023). We utilise XG-

Boot due to a combination of agricultural yield prediction, financial prediction and the use of both economic and environmental variables as XGBoost is known to perform well with mixed types of predictor domains (Yuanchao Li and Qin, 2024; Zhang et al., 2023). Furthermore we choose XGBoost as it has a good performance in predictions whislt allowing the use of directly comparable metrics to sanity check models against prior research (such as yield using R^2), offering insight into the relative performance of models lacking prior reference points in the literature such as revenue and operating costs and making the model more interpretable to audiences familiar with regression models (He et al., 2022; Laurent et al., 2021).

The ability to classify and predict continuous response variables and categoric variables alongside one another was also a consideration in the use of XGBoost, as both were contained in the data. XGBoost was also used due to its ability to handle sparse data, which was present within this dataset due to the voluntary nature of data entry, with many fields being left blank during data collection. Tree based methods also do not require data to be transformed prior to analyses; this consideration was taken into account so that specific partitions of values could be evaluated more easily and understood within original units of the data (D. Mariadass et al., 2022). A further consideration in its use was the level of interpretability offered through measures of 'relative importance' allowing for the ability to identify and rank variables and interactions by contribution to predictions (Chen and Guestrin, 2016).

An XGBoost model was trained for each variable so that every variable's relative importance could be caculated. This process was done three times using three iterations of data (three models for each variable). The first models were trained on the original SWA data set, the second were trained on a dataset that incorporated external data for each region and the final were trained on data with continuous variables transformed to be expressed as a ratio of vineyard area. The final dataset that consisted of ratios also included the extra regional data (but not in ratio form).

2.4. Sankey and Chord Diagrams

Originally created by Sankey to depict different pressures in steam engines (Yu and Silva, 2017) we leverage Sankey diagrams to illustrate the different impact or 'pressure' each variable has on one another through the use of measurements of variable importance. Sankey and Chord diagrams were constructed using the Holoviews python library (Rudiger et al., 2020). Sankey diagrams (depicted on the left as section A in figures) show the top 10 contributing factors to a variables prediction using XGBoost and Chord diagrams, a circular represntation of Sankey diagrams (depicted on the right as section B in figures) show how each of the top 10 factors relate to one another by measures of relative importance. Both Chord and Sankey diagrams illustrate variable importance through the size of the bands between two variables. The number at the end of a connection in the diagrams indicates a variabless importance (the number of times it appeared within the ensemble).

2.5. Variable Importance

XGBoost creates a large number of decision trees in the ensemble, it is hard to directly interpret the model and the derived intricate relationship between the variables. Variable importance can be measured in multiple ways, in this paper we used the frequency of a variable appearing as a node within the ensemble as a measure of its importance. This measure can be interpreted as how often a variable was the optimal choice in reducing the loss function of the ensemble. Multiclass variables are given an importance score for each individual class; for example, in the first set of analyses each specific region will have its own importance score, as will Year, Irrigation Type, etc (see Table 1).

2.6. Validation

The predictive accuracy of each tree was assessed through a validation process. For each model, a sample of 80% of the data was used for training the model and the remaining 20% was used for testing and validation. Categorical data were stratified to conserve the same proportion of class occurrences between the training, testing and validation data. The models were validated using 10 repetitions of the sampling process (10-fold cross validation). R^2 scores were used to determine the best regression models during validation. For analyses with continuous responses R^2 was used instead of RMSE to allow the comparison of models with different units to each other when considering how well each model extrapolated to further data. For binary and multiclass variables, validation was summarised through the accuracy, the proportion of true negatives and positives.

2.7. Hyperparameters

As part of the utilising the XGBoost model the hyperparameters of the model were tuned. The XGBoost library incorporates regularisation techniques built into the software to mitigate over-fitting and enhance model generalisation. This allowed us to utilise cross validated grid search functions when selecting for better performing hyperparameters. This method required three distinct types of metrics to be used for the three types of variables incorporated into the analysis (multiclass, binary and numeric). For consistency the metrics utilised by the grid search were aligned with the error functions used when training the model on those variables. The performance measure for model selection was root-mean-square error for continuous variables. The receiver operator characteristic's area under the curve was used for binary variables (Hanley and McNeil, 1982). And, multiclass variables utilised the one verse one approach to minimise sensitivity to class disparity (Ferri et al., 2009; Hand and Till, 2001).

3. Results

3.1. Revenue

There was little difference in predictive power between the inclusion of climatic and regional data and the use of region alone. With the use of extra regional data achieving an R^2 of 0.76 (with a standard deviation of 0.13), and when not using the extra regional data achieving an R^2 of 0.77 (with a standard deviation of 0.15). The higher number of variables included as part of the regional data likely resulting in a lower variance in predictive power when including the extra regional data.

The most notable difference was in the importance of the predictors, where elevation was a had a high relative importance of 8, only surpassed by diesel and yield, with 9 and 45 relative importance respectively (TODO: see figure). The lower variance is also reflected in the lower amount of nodes (or partitions) required when leveraging the extra regional parameters, with a reduction in hundreds of splits between the two models. Without the extra regional parameters it was found that fuel use (petrol 307 and diesel 144), yield (285), size (216) and water use (199) held the highest relative importance (TODO: see figure).

Even without the extra regional parameters, overall region contributed to 234 nodes in the ensemble making it collectively the third most important variable. This places equal importance within both models on region however the nature of regions contribution is more generalised using elevation when included as a parameter. This alone does not necessarily make elevation a direct contributing factor but links regions that of similar terrain together. The relevance of this was noted when reviewing regional missclassifications, where neighbouring regions were often misclassified as each other. The extra numerical parameters that can be partitioned likely gives the algorithm a greater ability to partition these regions together using fewer nodes.

3.2. Operating Costs

Compared to revenue, the predictive performance of the XGBoost model for operating cost was slightly better when not using the extra regional parameters, with an R^2 of 0.80 (with a standard deviation of 0.10). Similarly, when predictin without extra regional data the most important predictors of operating cost were fuel, water, area and yield (see figure 2).

A major difference was also in the poorer performance of the model with extra regional parameters. Although achieving a similar R^2 of 0.78 and a standard deviation of 0.12 the difference lay in an outlier model recording an R^2 of 0.08. This divergence likely being due to over generalising using

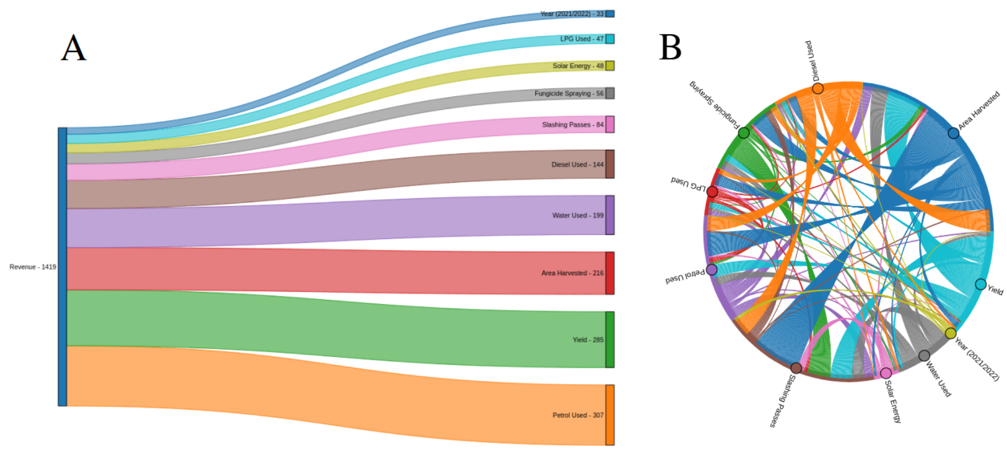

Figure 1: The left-hand side depicts the 10 most important variables in predicting revenue using XGBoost as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

elevation to match neighbouring regions. The removal of this radical leaves the inclusion of extra regional parameters outperforming the models without them, however it overlooks the major pitfall in the ease of misattributing factors and causality when predicting these variables.

A surprising difference between operating cost and revenue was the change in relative importance of activities involving tractor passes where the use of fungicide was more important for operational costs, compared to revenue, where slashing was more important (see Figure 3). This difference was only found when not including the extra regional parameters. The model including extra regional parameters reflected an identical hierarchy of importance to its revenue counter part (TODO: see figs).

The connection between spraying and operational costs is intuitive in that it utilises both the expense of equipment and resources. However, it is surprising that although spraying is considered important when extra regional paramters were not included, 'area not harvested due to disease' was not, even though disease would be a direct cause. The lack of importance on disease directly could be due to a low amount reported in the dataset (137 vineyards). The reason for spraying was also unfortunately not part of the data, and could be in response to a variety of factors such as other vineyards within the region having disease or preventative sprays. The variables that feed into these decisions are also very different with diesel having the highest relative importance to slashing, and area having the greatest relative importance to the need for fungicide.

Again, Region played a determining factor overall, contributing to 334 nodes within the ensemble making it the most important variable when considering all regions together. It was surprising that electricity, slashing and spraying passes were not more prominent in operating costs due to the intrinsic nature as an agricultural expense. However, a consideration for a bias within the dataset may be explanatory towards to the lack of these factors contributing to expenses, with the dataset being derived from vineyards actively participating within a sustainability program.

3.3. Region

Region was a highly informative variable based on measures of importance for both operating cost and revenue. As noted above, Region was the third most important variable for determining revenue. The Barossa Valley region and Tasmania were the two most important regions in relation to revenue; these two regions are considered to be some of the highest revenue per hectare

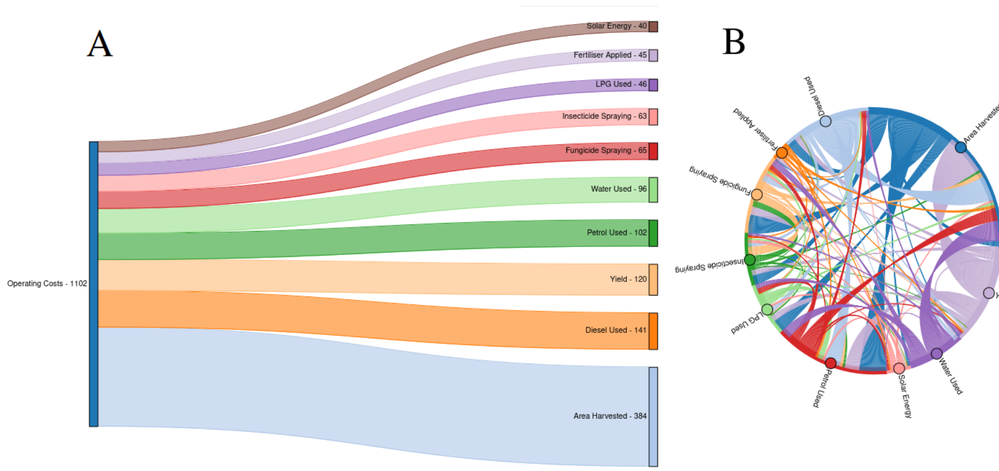

Figure 2: The left-hand side, A, depicts the 10 most relative important variables in predicting Operating Costs using XGBoost as a measure of node occurrence, using a Sankey diagram. The number at the end of each band in the diagram is that variable's importance. The right-hand side, B, depicts the importance of the 10 variables in Sankey diagram relative to one another.

regions in Australia (Wine Australia, 2022). These two regions are also relative opposites in winegrowing climates with the Barossa having a warm and dry climate focussing on Shiraz grapes and Tasmania having a cool wet climate that favours Pinot/Chardonnay (Wine Australia, 2022).

As also noted above, Region was also a key determinant of operating costs. Tasmania had the highest relative importance, followed by the Adelaide Hills. In contrast, the regions of the highest relative importance were warmer and drier, such as the Barossa. The higher relative importance of fungicide spraying is the likely due to fungal pressure being greater in cooler wetter regions variables than in drier regions.

The XGBoost ensemble for Region achieved an accuracy of 56.82% (and

50.58% validation accuracy). The difference in accuracy compared to the other models is in part due to the large number of classes (58 regions). The ensemble had an emphasis on area, water, fuel and yield as determining factors (see Figure (3).

A number of regions had lower reporting rates, resulting in much poorer classification performance. The regions with the most samples performed the best likely due to the disparity in sample sizes. Bordering regions were routinely grouped together and misclassified as the same region. When scrutinising each class explicitly, the two areas that effected the most from this were the Limestone Coast (cool coastal areas in South Australia) and the warmer inland regions along the Murray Darling.

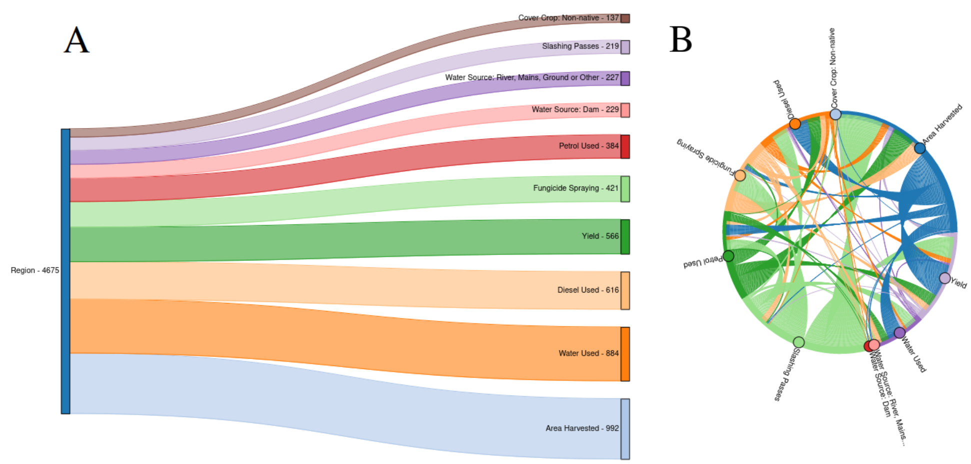

Figure 3: The left-hand side, A, depicts the 10 most relative important variables in predicting Region using XGBoost as a measure of node occurrence, using a Sankey diagram. The number at the end of each band in the diagram is that variable's importance. The right-hand side, B, depicts the importance of the 10 variables in Sankey diagram relative to one another.

3.4. Results per hectare

Both operating cost and revenue were predicted as ratios of area (revenue and operating cost per vineyard hectare). In both cases it was found that the models performed poorly with operating costs recording an R^2 of 0.24 (with a standard deviation of 0.15) and revenue an R^2 of 0.32 (with a standard deviation of 0.14).

4. Discussion

The significance of yield and region is demonstrated by their continual apparance as important variables when predicting operating costs and revenue. Several physical parameters such as climate, geology and soil are predetermined by a vineyard's location, making it a widely considered key determinant of grape yield and quality (Abbal et al., 2016; Agosta et al., 2012; Fraga et al., 2017). The relationships between vineyard resource use, operations and geographical properties is a complex one, illustrated through the highly interelated properties demonstrated in the chord diagrams. The difference in resources available between regions is well illustrated when considering the relative importance that water use had in predicting region. The difference in readily available resources between different regions is also easily demosntrated when observing the partitioning of river water as the primary source of water to identify vineyards located in the riverland (with 456 of 472 vineyards in the riverland utilising riverwater).

The availability of resources and geographical features of a region are significant but are also a potentially determining factor in the types of operational decisions made, and thus the reasoning as to variation in revenue and operating costs. This can be seen in the addition of extra regional paramters that add greater context to understanding the why or cause of operating costs and revenue. The effect of different operational consideration can be reflected in higher costs likely incurred in regions of greater variations in slope requiring more specialised equipment, or reflected in the types of resources available in some regions. The specifics of a vineyards' site are of incredible importance when determining what causes higher operational costs or revenue. Although it is useful to be able to predict and compare regions and their revenue and operational costs, a greater nuace to help understand the why behind these decisions, would help in specifically guiding operational decisions. AN example of this can is whether the reduction of tillage operations through optimising tractor efficiency would be useful and how to optimise tractor use for a specific operation. This example is chosen as while this practice is undertaken to reduce energy use in vineyards, decreasing running costs, as well as reducing soil compaction (Capello et al., 2019). The interrelatedness of this decisions is far reaching as increase in tractor use can cause soil compaction which has been shown to further increase water runoff (Capello et al., 2020). With runoff itself being a significant factor during extreme rain events which can lead to large scale soil deposition, creating further erosion and removing topsoil and having wide spread effects for a vineyard.

Further to the consideration of including specific operations is hindered in this model due to the sample being derived specifically from vineyards already within a sustainable program. Making the sample inherintly biased towards the use of sustainable practices. A keen example is the use of techniques such as cover crops. Cover crops are an example of a sustainable practice in viticulture in which the area between vine rows is seeded with a crop such as grasses or native vegetation. The primary reason for employing cover crops is to reduce the presence of disease and weeds (Delpuech and Metay, 2018). The benefit of reducing diseases and weeds is especially notable, as there is less cause to utilise heavy machinery for spraying herbicides and fungicides, or for mechanical weeding (Capello et al., 2019). The presence of a cover crop can also help to increase soil water retention, reducing erosion and water runoff in shallow soils, having been shown to mitigate runoff during rain events by over 65% (Capello et al., 2020). However, cover crops can introduce competition with grapevines and may reduce yield depending upon the plants used and the density of the cover crop (Capello et al., 2019; Delpuech and Metay, 2018; Gosling and Shepherd, 2005; Monteiro and Lopes, 2007). A coverage of only 30% is required to provide protection against erosion, yet increased cover provides the benefits of greater biodiversity at the risk of yield (Delpuech and Metay, 2018). The presence of cover crops within the sample is reflects this bias, where just over 85% (5272) of vineyards utilised some form of cover crop such as grassing and only just under 4% (225) used only bare soil (with the remaining 552 utilising a combination). The high percentage of vineyards using this type of sustainable practice means that its effect is will not be prominent within the model, and can only show what practices would further improve those already implementing these techniques, and how they are connected to these operating costs. A strength of utilising XGBoost in this context is that, a subset of particular interest can be leveraged to focus in on the combination of factors that would contribute to the specific concieved scenario.

Warmer regions are known to be beneficial in hastening the ripening process of winegrapes (Webb et al., 2011). Warmer regions are also associated with lower quality grapes, caused largely due to this hastened ripening (Botting et al., 1996). It is likely that the combination of larger vineyards with higher water use is a determining factor in classifying regions which favour larger production of grapes; reflected through region using water use so prominently in the XGBoost ensemble. The link to water resources in defining regions is also an important consideration, as vineyards can leverage higher irrigation rates if water resources are available. A further consideration in the link between revenue and region is that grape prices are set at a regional level by buyers (Wine Australia, 2022). It is also important to consider that some regions carry particular fame regarding the quality of their produce such as Tasmania, the Hunter Valley and Barossa Valley (Halliday, 2009). This classification can be contrasted with other warmer regions of higher rainfall that use the warmer climate to concentrate their grapes, increasing the flavour profile (Goodwin I, Jerie P, 1992; MG McCarthy et al., 1986).

Yield is sometimes restricted simply through access to water resources. Regions are likely to have varying access to different water sources, such as those along the River Murray being able to utilise river water for crops, unlike most coastal regions which may be drawing from surface or underground water sources. Similarly, the connection between region and fuel use is likely an indicator of the level of infrastructure within the region due to vineyards in regions without pressurised water needing to use fuel or electricity to pressurise their irrigation systems. Although infrastructure between regions, especially further from cities is likely to vary, fuel price itself has little variation across regional Australia. It is reported by the Australian Competition and Consumer Commission that during the period of this data, that regional fuel prices tended to be higher (+5.4c/litre) and more stable than urban prices due to their primary driver being international market trends (AIP, 2019). The importance between fuel and other variables is a complicated interaction. The size, number of blocks, types and age of equipment will contribute to the efficiency of its use and the amount required across a site. It is likely that larger operations will generaly gain from economies of scale but also risk further incurring costs from the need to redeploy equipment. A further connection between region and fuel is the possible requirement of more specialty equipment, either due to regional practices differing or physical requirements such as greater inclines. However, the style of management will also greatly contribute to how efficient both fuel and water are used, which is difficult to account for through the use of a metric.

Operational costs showed similar importance across fuel, water and tractor use. The dominating factor of area likely played a large part in determining how costly a tractor pass would be, or in defining the ratio of water applied to the amount of vines. The relative importance was high for area but much lower in general across the other variables, which could indicate the need to be specific when attempting to determine the cause of a operational cost. Although these analyses attempted to capture the complexity between how variables interacted when determining operational costs (see Figure 2), in reality these relationships are likely even more complicated. An example of how interrelated operational costs can be, is the optimisation of tractor passes to achieve multiple goals in a pass, being shown to reduce energy use in vineyards, decreasing running costs, as well as reducing soil compaction (Capello et al., 2019).

When determining revenue, similar variables were used to operational cost; with region also being of high variable importance relative to other variables (when considering all regions together in importance). It is difficult to extrapolate the specific influence of location on a vineyard's outcomes due to the broad and varying definition of a region. Utilising the Geographical Indicator regions defined by Wine Australia (Australia, 2021b) is a limitation in one way, as it is too broad to fully capture a vineyards location and how that influences variables at a more granular level. However, as buyers set prices at regional levels, it is still important to consider this factor.

Decisions made on the ground have far-reaching effects and are difficult to completely capture. A larger number of tractor passes used as a preventative measure for occurrences such as disease may incur higher operational costs but could be critical in preventing long term losses. Although the models demonstrated a good predictive fit, the ability to predict operational costs is limited by the variables incorporated in the analysis. Other factors such as erosion and soil health are also influenced by tractor use and would contribute to these operational costs but are difficult to measure and were not available as part of the data (Capello et al., 2019, 2020). The data collection process being voluntary and part of a sustainable program also limitted the ability to compare what happened between those who had to abandon crops due to disease, pests or other catastrophes such as fire, in part due to a lack of incentive to record as part of the SWA program. Furthermore, no comparison can be made between those that have chosen to mothball as a response to predicted outcomes, or external presssures due to them not being part of the data. Although this dataset contained vineyards that suffered partial losses due to disease, these limitations offer an avenue for further study that could benefit decision processes and variable relevance regarding mothballing, crop loss and external pressures. Without fully capturing more granular activities, for example the specific of tractor operastions and their differing fuel consumptions, it is difficult to determine what decisions specifically influence the operational costs. Reductions in fuel, water and tractor use are obvious methods to reduce operational costs but not necessarily achievable decisions when considering external risks such as disease.

The reasoning for any particular decision can be widely varying. More sophisticated models, specifically those that utilise expert opinion, may also help to capture and address the decision-making process. An example is the optimisation of fungicide sprays using Bayesian models that forecast disease risk (Lu et al., 2020).

Separately, revenue and operating cost did have a greater predictability than their counterpart profit (see Appendix). The disparity in accuracy between profit and other economic outcomes is reflective of the complexity in trying to address challenges such as climate change, disease and changing market demands (Wine Australia, 2020, 2021, 2022). The difference between turning a profit or loss is dependent on predictable factors unforecasted factors, farming practice and farmers' decisions. The difference between vineyards that make profit and those that do not could be a multitude of factors including differences in farming practices not captured within this study.

5. Conclusion

This study has provided valuable insights into the multifaceted dynamics governing operational costs and revenue in vineyards. The impact of different regions highlighted the complex interrelatedness of variables within a vineyard. We relate how factors such as water and fuel intersect to impact operational costs and how different seasonal events affect these operations; as well as the significance of context-specific decision-making. While this investigation utilised a broad regional classification, the potential benefits of adopting a more nuanced approach and incorporating expert knowledge have been highlighted. Further work could pursue causal models and the creation of decision support systems. It is difficult to untangle the predictive and correlative nature of a variable compared to the causal reasons. By delving deeper into the complex interplay of variables, further advancements can be made in optimising vineyard management strategies for lowering operational costs, increasing revenue and enhancing sustainability.

References

Abad, J., Hermoso de Mendoza, I., Marín, D., Orcaray, L., Santesteban, L.G., 2021. Cover Crops in Viticulture. A Systematic Review (1):
Implications on Soil Characteristics and Biodiversity in Vineyard. OENO One 55, 295–312. doi:10.20870/oeno-one.2021.55.1.3599.

Abbal, P., Sablayrolles, J.M., Matzner-Lober, É., Boursiquot, J.M., Baudrit, C., Carbonneau, A., 2016. Decision Support System for Vine Growers

Based on a Bayesian Network. Journal of agricultural, biological, and environmental statistics 21, 131–151. doi:10.1007/s13253-015-0233-2.

Agosta, E., Canziani, P., Cavagnaro, M., 2012. Regional Climate Variability Impacts on the Annual Grape Yield in Mendoza, Argentina. Journal of Applied Meteorology and Climatology 51, 993–1009.

AIP, 2019. Facts About Prices in Regional and Country Areas.

Attorney-General's Department, 2010. Wine Australia Corporation Act 1980.

Australia, W., 2021a. Australian Wine: Production, Sales and Inventory 2019– 20.

Australia, W., 2021b. Wine Australia-Open Data.

Baiano, A., 2021. An Overview on Sustainability in the Wine Production Chain. Beverages 7. doi:10.3390/beverages7010015.

Botting, D., Dry, P., Iland, P., 1996. Canopy Architecture-Implications for Shiraz Grown in a Hot, Arid Climate .

Capello, G., Biddoccu, M., Cavallo, E., 2020. Permanent Cover for Soil and Water Conservation in Mechanized Vineyards: A Study Case in Piedmont, NW Italy 15.

Capello, G., Biddoccu, M., Ferraris, S., Cavallo, E., 2019. Effects of Tractor Passes on Hydrological and Soil Erosion Processes in Tilled and Grassed Vineyards. Water 11. doi:10.3390/w11102118.

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, ACM, New York, NY, USA. pp. 785–794. doi:10.1145/2939672.2939785.

Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Using Data Mining for Wine Quality Assessment, in: Discovery Science: 12th International Conference, DS 2009, Porto, Portugal, October 3-5, 2009 12, Springer. pp. 66–79.

D. Mariadass, E. G. Moung, M. M. Sufian, A. Farzamnia, 2022. Extreme Gradient Boosting (XGBoost) Regressor and Shapley Additive Explanation for Crop Yield Prediction in Agriculture, in: 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), pp. 219–224. doi:10.1109/ICCKE57176.2022.9960069.

Ferri, C., Hernández-Orallo, J., Modroiu, R., 2009. An Experimental Comparison of Performance Measures for Classification. Pattern Recognition Letters 30, 27–38. doi:10.1016/j.patrec.2008.08.010.

Fraga, H., Costa, R., Santos, J.A., 2017. Multivariate Clustering of Viticultural Terroirs in the Douro Winemaking Region. Ciência Téc. Vitiv. 32, 142–153.

G. van Rossum, 1995. Python Tutorial, Technical Report CS-R9526.

Goodwin I, Jerie P, 1992. Regulated Deficit Irrigation: Concept to Practice. Advances in Vineyard Irrigation. Australian and New Zealand Wine Industry Journal 7.

Hall, A., Lamb, D.W., Holzapfel, B.P., Louis, J.P., 2011. Within-Season Temporal Variation in Correlations between Vineyard Canopy and Winegrape Composition and Yield. Precision Agriculture 12, 103–117.

Halliday, J.C.J.C., 2009. Australian Wine Encyclopedia. Hardie Grant Books, VIC.

Hand, D.J., Till, R.J., 2001. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 45, 171–186. doi:10.1023/A:1010920819831.

Hanley, J.A., McNeil, B.J., 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. Radiology 143, 29–36.

He, L., Fang, W., Zhao, G., Wu, Z., Fu, L., Li, R., Majeed, Y., Dhupia, J., 2022. Fruit Yield Prediction and Estimation in Orchards: A State-of-the-Art Comprehensive Review for Both Direct and Indirect Methods. Computers and Electronics in Agriculture 195, 106812. doi:10.1016/j.compag.2022.106812.

Kasimati, A., Espejo-García, B., Darra, N., Fountas, S., 2022. Predicting Grape Sugar Content under Quality Attributes Using Normalized Difference Vegetation Index Data and Automated Machine Learning. Sensors 22. doi:10.3390/s22093249.

Kisten, M., Ezugwu, A., Olusanya, M., 2024. Explainable artificial intelligence model for predictive maintenance in smart agricultural facilities. IEEE access : practical innovations, open solutions 12, 1–20. doi:10.1109/ACCESS.2024.3365586.

Laurent, C., Oger, B., Taylor, J.A., Scholasch, T., Metay, A., Tisseyre, B., 2021. A Review of the Issues, Methods and Perspectives for Yield Estimation, Prediction and Forecasting in Viticulture. European Journal of Agronomy 130, 126339. doi:10.1016/j.eja.2021.126339.

Li, Y., Zeng, H., Zhang, M., Wu, B., Qin, X., 2024. Global de-trending significantly improves the accuracy of XGBoost-based county-level maize and soybean yield prediction in the Midwestern United States. GIScience & Remote Sensing 61, 2349341. doi:10.1080/15481603.2024.2349341.

Lu, W., Newlands, N.K., Carisse, O., Atkinson, D.E., Cannon, A.J., 2020. Disease Risk Forecasting with Bayesian Learning Networks: Application to Grape Powdery Mildew (Erysiphe Necator) in Vineyards. Agronomy (Basel) 10, 622. doi:10.3390/agronomy10050622.

Luke Mancini, 2020. Understanding the Australian Wine Industry: A Growers Guide to the Background and Participants of the Wine Grape Industry.

Mariani, A., Vastola, A., 2015. Sustainable Winegrowing: Current Perspectives. International Journal of Wine Research 7, 37–48.

MG McCarthy, RM Cirami, DG Furkaliev, 1986. The Effect of Crop Load and Vegetative Growth Control on Wine Quality. .

Montalvo-Falcón, J.V., Sánchez-García, E., Marco-Lajara, B., Martínez-Falcó, J., 2023. Sustainability Research in the Wine Industry: A Bibliometric Approach. Agronomy 13. doi:10.3390/agronomy13030871.

OECD, 2019. Innovation, Productivity and Sustainability in Food and Agriculture.

Oliver, D., Bramley, R., Riches, D., Porter, I., Edwards, J., 2013. Review: Soil Physical and Chemical Properties as Indicators of Soil Quality in Australian Viticulture. Australian Journal of Grape and Wine Research 19, 129–139. doi:10.1111/ajgw.12016.

Ravi, R., Baranidharan, D., 2020. Crop yield prediction using XG boost algorithm. International Journal of Recent Technology and Engineering (IJRTE) 8, 3516–3520. doi:10.35940/ijrte.D9547.018520.

Rudiger, P., Stevens, J.L., Bednar, J.A., Nijholt, B., Andrew, B, C., Randelhoff, A., Mease, J., Tenner, V., maxalbert, Kaiser, M., ea42gh, Samuels, J., stonebig, LB, F., Tolmie, A., Stephan, D., Lowe, S., Bampton, J., henriqueribeiro, Lustig, I., Signell, J., Bois, J., Talirz, L., Barth, L., Liquet, M., Rachum, R., Langer, Y., arabidopsis, kbowen, 2020. Holoviz/Holoviews: Version 1.13.3. doi:10.5281/zenodo.3904606.

SOAR, C., SADRAS, V., PETRIE, P., 2008. Climate Drivers of Red Wine Quality in Four Contrasting Australian Wine Regions. Australian journal of grape and wine research 14, 78–90. doi:10.1111/j.1755-0238.2008.00011.x.

Srivastava, S., Sadistap, S., 2018. Non-Destructive Sensing Methods for Quality Assessment of on-Tree Fruits: A Review. Journal of Food Measurement and Characterization 12, 497–526.

SWA, S.W.A., 2022. Sustainable Wingrowing Australia.

Webb, L.B., Whetton, P.H., Barlow, E.W.R., 2011. Observed Trends in

Winegrape Maturity in Australia. Global change biology 17, 2707–2719. doi:10.1111/j.1365-2486.2011.02434.x.

Wine Australia, 2020. National Vintage Report 2020 .

Wine Australia, 2021. National Vintage Report 2021 .

Wine Australia, 2022. National Vintage Report 2022 .

Yu, B., Silva, C.T., 2017. VisFlow - web-based visualization framework for tabular data with a subset flow model. IEEE Transactions on Visualization and Computer Graphics 23, 251–260. doi:10.1109/TVCG.2016.2598497.

Yuanchao Li, Hongwei Zeng, M.Z., Qin, X., 2024. Global de-trending significantly improves the accuracy of XGBoost-based county-level maize and soybean yield prediction in the Midwestern United States. GIScience & Remote Sensing 61, 2349341. doi:10.1080/15481603.2024.2349341, arXiv:https://doi.org/10.1080/15481603.2024.2349341.

Zhang, T., Zhu, W., Wu, Y., Wu, Z., Zhang, C., Hu, X., 2023. An explainable financial risk early warning model based on the DS-XGBoost model. Finance Research Letters 56, 104045. doi:10.1016/j.frl.2023.104045.

Appendix A. Continuous variables

Table A.2 below shows the ranges of each of the continuous variables:

Table A.2: Summary statistics of continuous variables used in XGBoost models.

	count	mean	std	min	0.25	0.5	0.75	max
Vineyard Solar	622	22916.89	104808	1	1170.75	5500	14866.25	2300000
Biodiesel	25	6635.932	11768.832104	1	200	500	10000	37216
Fungicide Spray	2260	7.724801	3.279794	1	6	7	9	68
LPG	958	327.831399	861.538804	1	40	95.835	240	11950
Petrol	4309	825.276809	1556.621119	1	135	306.66	903	38568
Insecticide Spray	1092	1.707189	1.316042	0	1	1	2	12
Water Used	5846	7301838	558206600	0.0007	13.2655	43	146.875	42680000000
Fertiliser	795	91149.89	483913.4	1	560	4759.5	45148.5	11358000
Diesel	5585	11677.070183	24380.588742	0.1267	1240	3850	12500	591000
Yield	5935	772.902449	2175.113895	0.03	68	192.3	601.8795	72305
Herbicide Spray	2026	2.646199	2.598899	0	2	2	3	103
Slashing	2290	3.311485	1.826788	1	2	3	4	26
Electricity	1014	58223.07	177626.3	0.019	2160	9637	36498.25	3000000
Area Harvested	6049	66.52604	133.4525	2.220446E-16	10.13	24.5	66.8	2436.15
Grape Revenue	875	377972	606286.8	1	76000	172964	386747	5700000
Operating Costs	853	314187.1	511522.6	1	57315	140000	327408	4482828

Appendix B. Categorical Variables

The tables below describe each possible class a multiclass variable could have taken and the frequency that it occured.

Appendix B.1. Water Source Types

Table B.3 below shows the different class types for water sources used by vineyards and their frequency of occurrences.

Table B.3: Frequency and class types of water types used by vineyards.

Water types	**frequency**
river water	1578
groundwater	1433
surface water dam	617
recycled water from other source	386
groundwater and surface water dam	256
not listed	235
mains water	170
river water and groundwater	147
groundwater and recycled water from other source	145
other water	101
river water and surface water dam	92

Continued on next page

Table B.3 – continued from previous page

Water types	frequency
groundwater and water applied for frost control	90
groundwater and mains water	76
river water and groundwater and surface water dam	70
recycled water from other source and mains water	63
groundwater and recycled water from other source and mains water	60
river water and mains water	57
surface water dam and mains water	56
groundwater and other water	33
river water and groundwater and mains water	30
groundwater and surface water dam and recycled water from other source	27
river water and water applied for frost control	27
groundwater and surface water dam and mains water	22
surface water dam and recycled water from other source	21

Continued on next page

Table B.3 – continued from previous page

Water types	frequency
river water and recycled water from other source	19
river water and other water	19
river water and surface water dam and mains water	18
river water and groundwater and surface water dam and mains water	18
mains water and other water	16
groundwater and surface water dam and water applied for frost control	12
surface water dam and other water	12
groundwater and recycled water from other source and other water	11
groundwater and surface water dam and recycled water from other source and mains water	8
recycled water from other source and mains water and other water	8
river water and recycled water from other source and mains water	8
river water and surface water dam and recycled water from other source	8

Continued on next page

Table B.3 – continued from previous page

Water types	**frequency**
surface water dam and mains water and other water	7
recycled water from other source and other water	7
river water and groundwater and recycled water from other source	6
groundwater and mains water and other water	5
groundwater and surface water dam and other water	5
groundwater and surface water dam and mains water and other water	5
river water and groundwater and recycled water from other source and mains water	5
river water and groundwater and water applied for frost control	5
river water and surface water dam and water applied for frost control	4
surface water dam and water applied for frost control	4

Continued on next page

Table B.3 – continued from previous page

Water types	**frequency**
river water and groundwater and surface water dam and recycled water from other source and mains water and other water	4
river water and groundwater and recycled water from other source and other water	3
groundwater and surface water dam and recycled water from other source and water applied for frost control	3
river water and groundwater and surface water dam and recycled water from other source	3
river water and recycled water from other source and other water	3
surface water dam and recycled water from other source and mains water	2
river water and recycled water from other source and mains water and water applied for frost control	2

Continued on next page

Table B.3 – continued from previous page

Water types	**frequency**
groundwater and surface water dam and recycled water from other source and mains water and other water	2
river water and groundwater and mains water and other water	2
river water and groundwater and surface water dam and other water	2
river water and surface water dam and other water	2
river water and mains water and water applied for frost control	2
river water and groundwater and surface water dam and recycled water from other source and mains water	2
river water and mains water and other water	2
river water and surface water dam and mains water and other water	2
river water and groundwater and mains water and water applied for frost control	1

Continued on next page

Table B.3 – continued from previous page

Water types	frequency
surface water dam and other water and water applied for frost control	1
water applied for frost control	1
groundwater and other water and water applied for frost control	1
other water and water applied for frost control	1
groundwater and surface water dam and recycled water from other source and other water and water applied for frost control	1
mains water and water applied for frost control	1
groundwater and surface water dam and recycled water from other source and other water	1
groundwater and mains water and water applied for frost control	1
river water and groundwater and surface water dam and mains water and other water	1

Continued on next page

Table B.3 – continued from previous page

Water types	frequency
river water and surface water dam and recycled water from other source and mains water	1

Appendix B.2. Cover Crop Types

Table B.4 below shows the different cover crop types used together and their frequency.

Table B.4: Frequency and class types of cover crop types used by vineyards.

Cover crop types	**frequency**
Cover crop types	frequency
permanent cover crop volunteer sward	1822
permanent cover crop non native	936
permanent cover crop native	490
annual cover crop	479
groundwater and surface water dam	406
annual cover crop and permanent cover crop volunteer sward	309
bare soil	225
permanent cover crop non native and permanent cover crop volunteer sward	214
annual cover crop and permanent cover crop non native	169
bare soil and permanent cover crop volunteer sward	129

Continued on next page

Table B.4 – continued from previous page

Cover crop types	**frequency**
bare soil and permanent cover crop non native	115
annual cover crop and permanent cover crop non native and permanent cover crop volunteer sward	101
bare soil and annual cover crop	93
permanent cover crop native and permanent cover crop volunteer sward	80
bare soil and permanent cover crop native	78
annual cover crop and permanent cover crop native	78
permanent cover crop native and permanent cover crop non native	68
permanent cover crop native and permanent cover crop non native and permanent cover crop volunteer sward	44
annual cover crop and permanent cover crop native and permanent cover crop non native and permanent cover crop volunteer sward	44

Continued on next page

Table B.4 – continued from previous page

Cover crop types	**frequency**
bare soil and annual cover crop and permanent cover crop volunteer sward	33
bare soil and permanent cover crop non native and permanent cover crop volunteer sward	26
annual cover crop and permanent cover crop native and permanent cover crop volunteer sward	17
bare soil and annual cover crop and permanent cover crop native	15
annual cover crop and permanent cover crop native and permanent cover crop non native	15
bare soil and annual cover crop and permanent cover crop non native	13
bare soil and annual cover crop and permanent cover crop native and permanent cover crop non native and permanent cover crop volunteer sward	12
bare soil and annual cover crop and permanent cover crop non native and permanent cover crop volunteer sward	11

Continued on next page

Table B.4 – continued from previous page

Cover crop types	**frequency**
bare soil and annual cover crop and permanent cover crop native and permanent cover crop non native	8
bare soil and permanent cover crop native and permanent cover crop non native	7
bare soil and permanent cover crop native and permanent cover crop volunteer sward	6
bare soil and permanent cover crop native and permanent cover crop non native and permanent cover crop volunteer sward	4
bare soil and annual cover crop and permanent cover crop native and permanent cover crop volunteer sward and	2

Appendix B.3. Irrigation Types

Below in Table B.5 are the frequency and different irrigation types.

Table B.5: Frequency and class types of irrigation types used by vineyards.

Irrigation types	**frequency**
Irrigation type	frequency
dripper	4800
dripper and non irrigated	342
Not listed	319
dripper and overhead sprinkler	201
dripper and undervine sprinkler	91
non irrigated	65
undervine sprinkler	53
dripper and flood	53
overhead sprinkler	46
dripper and overhead sprinkler and undervine sprinkler	28
overhead sprinkler and undervine sprinkler	12
dripper and non irrigated and overhead sprinkler	11
flood and undervine sprinkler	10

Continued on next page

Table B.5 – continued from previous page

Irrigation types	frequency
dripper and flood and undervine sprinkler	7
dripper and flood and non irrigated and overhead sprinkler and undervine sprinkler	3
dripper and flood and overhead sprinkler	3
non irrigated and undervine sprinkler	2
dripper and flood and non irrigated	1
dripper and non irrigated and overhead sprinkler and undervine sprinkler	1
flood and	1

Appendix B.4. Irrigation Energy Type

Below, Table B.6 shows the different types of energy used to power vineyards and their frequency.

Table B.6: Frequency and class types of irrigation energy types used by vineyards.

Irrigation Energy types	**frequency**
Irrigation energy type	frequency
electricity	2162
not listed	2053
pressure	586
electricity and pressure	396
diesel	254
diesel and electricity	227
electricity and solar	96
diesel and electricity and pressure	90
diesel and pressure	74
solar	50
electricity and pressure and solar	23
diesel and electricity and solar	14
diesel and electricity and pressure and solar	10
pressure and solar	9

Continued on next page

Table B.6 – continued from previous page

Irrigation Energy types	frequency
diesel and solar	4
diesel and pressure and solar and	1

Appendix B.5. Year

Below in Table B.7 is the list of years and the number of sample collected in each.

Table B.7: Frequency and class types of year

Year	**frequency**
Year	frequency
2021/2022	954
2020/2021	860
2019/2020	599
2012/2013	590
2013/2014	549
2015/2016	548
2014/2015	505
2017/2018	493
2016/2017	485
2018/2019	466

Appendix B.6. Region

Below in Table B.8 are the number of collected samples for each region.

Table B.8: Frequency and class types of regions.

Regions	**frequency**
giregion	frequency
McLaren Vale	1195
Barossa Valley	584
Murray Darling	521
Riverland	472
Adelaide Hills	454
Langhorne Creek	347
Margaret River	344
Coonawarra	284
Padthaway	202
Wrattonbully	195
Clare Valley	149
Yarra Valley	122
Eden Valley	92
Tasmania	89
Swan Hill	83
Grampians	73
Orange	72

Continued on next page

Table B.8 – continued from previous page

Regions	**frequency**
Hunter Valley	70
Bendigo	53
Great Southern	51
Rutherglen	41
Robe	36
Tumbarumba	35
Mornington Peninsula	32
King Valley	32
Southern Fleurieu	30
Heathcote	29
Adelaide Plains	25
Currency Creek	24
	23
Henty	22
Canberra District	21
Southern Flinders Ranges	20
Upper Goulburn	20
Mudgee	20
Mount Benson	20
Other	19
Riverina	18
Alpine Valleys	15

Continued on next page

Table B.8 – continued from previous page

Regions	**frequency**
Barossa Zone	14
Pemberton	12
Mount Gambier	11
Blackwood Valley	10
Kangaroo Island	10
Big Rivers Zone Other	9
Geographe	7
Cowra	6
Gundagai	5
Strathbogie Ranges	5
Glenrowan	4
Geelong	4
Swan District	4
Goulburn Valley	3
Beechworth	3
Southern Highlands	3
Macedon Ranges	2
Pyrenees	2
Sunbury	1

Appendix C. XGBoost

Following Chen and Guestrin (Chen and Guestrin, 2016), XGBoost predicted a value y_i from the input x_i. The method of prediction is achieved through a tree ensemble model, using K additive functions to predict the output. Each of f_k functions is a classification or regression tree, such that all functions are in the set of all decision trees, given by $\mathcal{F}$, is defined by $f(x) = \omega_{q(x)}(q : \mathbb{R}^m \to T, \omega \in \mathbb{R}^T)$. Where each function corresponds to an independent tree structure q of ω weights. Each tree has T leaves, which contain a continuous score, represented by ω_i for the i-th leaf. The final prediction is determined by the sum of the score of the corresponding leaves, given by:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), f_K \in \mathcal{F}, \tag{C.1}$$

The set of functions, $\mathcal{F}$, used by the tree is determined by minimising a regularised objective function, $\mathcal{L}$ given by:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i^{t-1} + f_t(x_i)) + \sum_k \Omega(f_K). \tag{C.2}$$

, where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||\omega||^2 \tag{C.3}$$

As predictions are made using additive tree functions, XGboost can be used for classification or regression. The difference between a prediction, $\phi(x_i)$, and actual variable, $f_k(x_i)$, is a differentiable convex loss function l. These properties of l allow the function to be versatile in which objective we choose to optimise for, which is also important in being able to process both continuous and categorical variables. To optimise l, the difference is calculated for the i-th instance at the t-th iteration.

Appendix C.1. Loss functions

The functions included as parameters in equation C.2 mean that traditional opimisation methods for Euclidean space cannot be used. Chen and Guestrin (Chen and Guestrin, 2016) illustrate, using Taylor expansions, that for a fixed structure $q(x)$ the optimal weight ω_j^* for a leaf j can be derived. Importantly a loss function can be used to fit a model iteratively to data. For this analysis several loss functions were used, as variables took the form of continuous, binary and multi-call data. The loss function for making a split within the tree structure is given by:

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma. \tag{C.4}$$

The tree structure being defined using left I_L and right I_R instance sets of nodes, with $I = I_L \cup I_R$. Instead of enumerating all possible tree structures, a greedy algorithm iteratively adds branches to the tree minimising $\mathcal{L}_{split}$ in (C.4). The frequency of a variable's occurrence within a tree is directly attributed to the minimisation of the loss function through the minimisation of $\mathcal{L}_{split}$.

The loss functions used for this analysis were the root-mean-square function for continuous variables, the logistic loss function for binary class variables, and the soft max function for Multiclass variables. All objective functions are defined within the SKlearn library (?), which was utilised via an API to the XGBoost library (Chen and Guestrin, 2016).

Appendix C.2. Year

The classification tree and XGBoost performed similarly for classifying year with 35.20% (6.28% standard deviation) and 51.81% (42.20% validation accuracy) respectively. Electricity and the type of irrigation were highly influential within the classification tree. Similarly, electricity was the most frequently occurring node in the XGBoost ensemble. Other variables such as slashing passes, and fungicide and herbicide spraying were more prevalent than in the classification tree. Weed and disease outbreaks are likely an influential factor when classifying different years, making the decisions to spray and slash unique factors that differ year to year. Climatic differences between years are likely tied to the influence of yield and water use.

Over half of the interrelated importance of the predictor variables is dominated by area harvested, yield and slashing passes. Although all the predictor variables are highly connected, their relative importance is not as prominent as the three major variables. It is of particular note of the relative importance of slashing passes to area, fuel and yield; as these are not directly related activities. The connection between the number of slashing and spraying passes is that those who do a set number of spraying or slashing passes tended to do that many passes for all slashing and spraying activities.

Appendix C.3. Profit

Predictions of profit performed poorly compared to operating cost and profit with an average R^2 of 0.2535 and standard deviation of 0.3126. With the large standard deviation being indicative of how unstable the models created were.

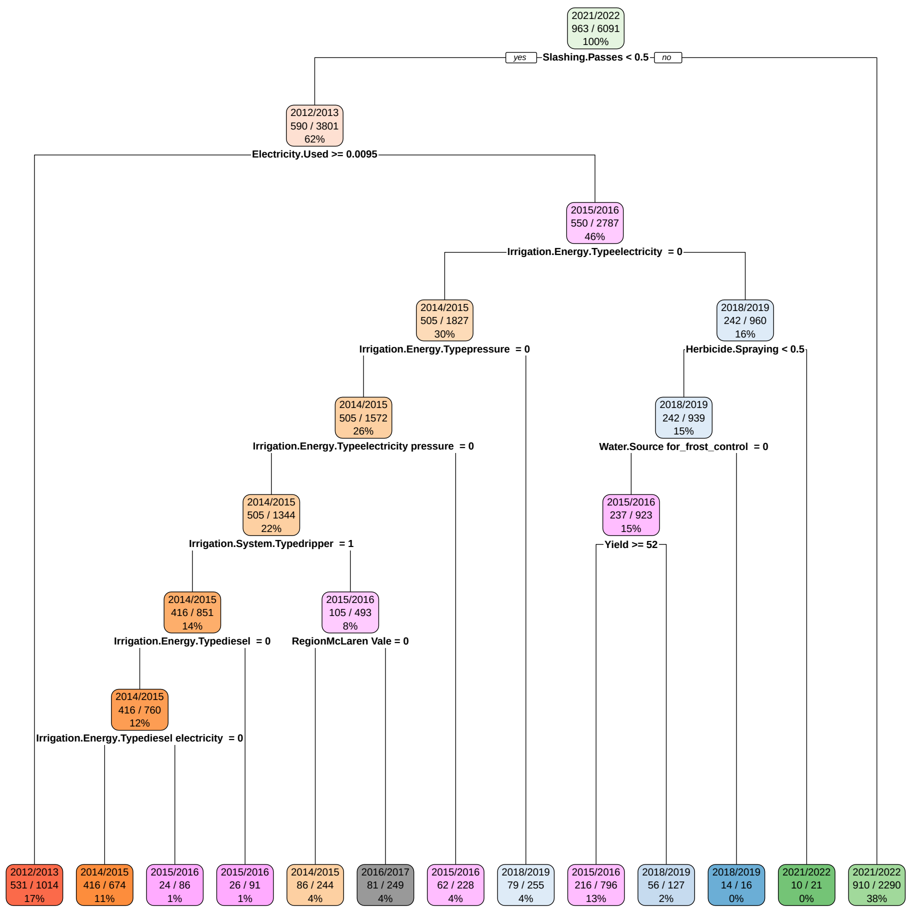

Figure C.4: Decision tree predicting Year. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

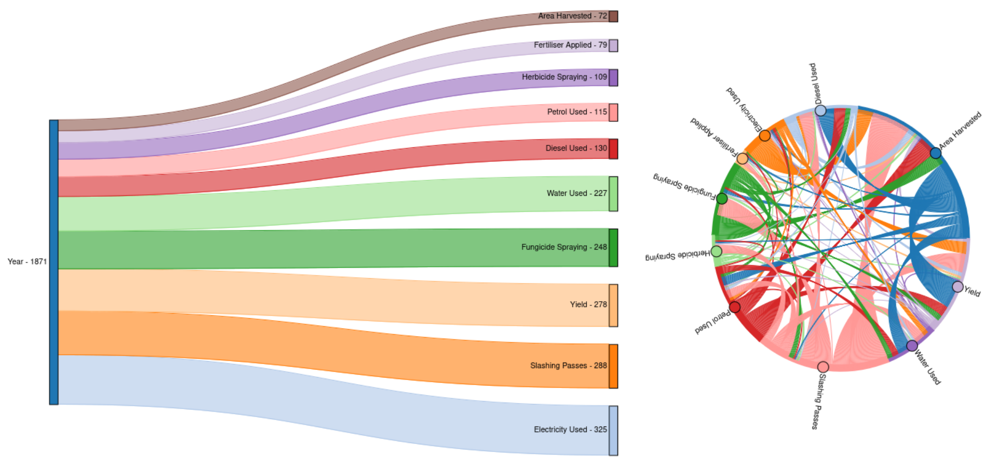

Figure C.5: The left-hand side depicts the 10 most relative important variables in predicting Year using XGBoost as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

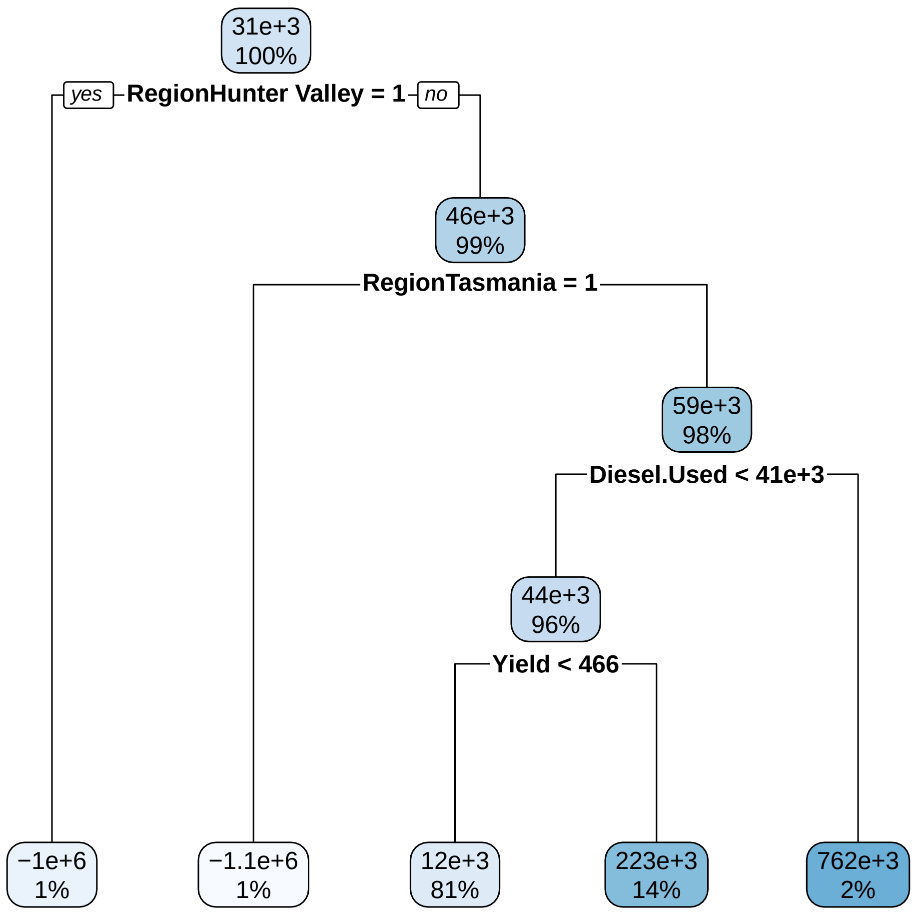

Figure C.6: Decision tree predicting revenue. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

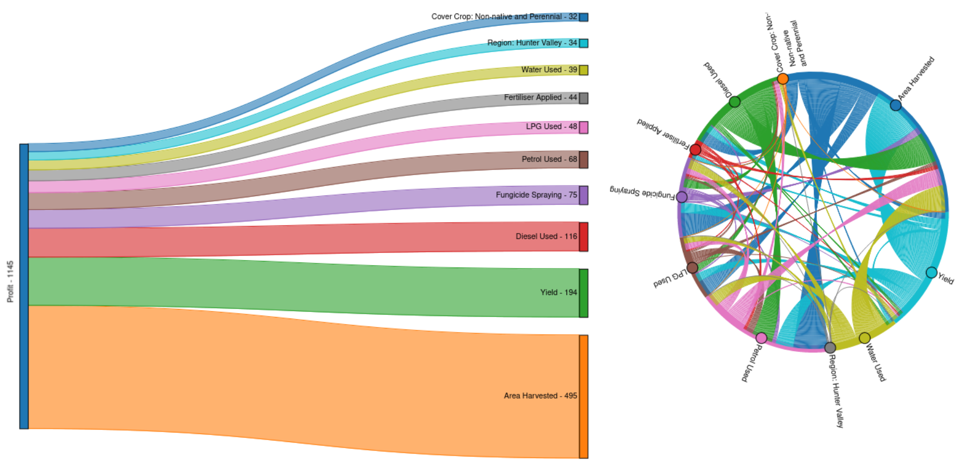

Figure C.7: The left-hand side depicts the 10 most relative important variables in predicting revenue using XGBoost as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.