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

Author[1,1,1]

1. Introduction

Historically strong demands for Australian wine have 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 has 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 are key to making data-informed decisions to aid in increasing a nations 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, where these relationships can lead to determining better and efficient methods and develop benchmarks with local growers (Luke Mancini, 2020).

An unprecedented amount of data regarding the Australian winegrowing industry has been collected through Sustainable Winegrowing Australia, offering new insights into the driving economic forces of the Australian wine industry. This dataset allows insights into the economic outcome of vineyards through the incorporation of operating costs and grape revenue from grape sales within the data. We use this data to study these economic outcomes and their statistical relationships to vineyards' utilisation of the resources. We further compare the relationships between different resources to address the extensive collinearity found within the data (Chen and Guestrin, 2016). We adopt XGBoosted models for this analysis because they are able to overcome multicollinearity as well as highlight the level of importance that predictor variables have on response variables.

2. Methods

2.1. Data

Data used in this analysis were obtained from Sustainable Winegrowing Australia. Australia's national wine industry sustainability program. The program aims to support grape-growers and winemakers in demonstrating and improving their sustainability (SWA, 2022). Data recorded by SWA are entered manually by winegrowers using a web based interface tool. A total of 6049 observations were collected from 2012/2013 to 2021/2022 financial years, with each observation comprising 23 variables reflecting a vineyard's state for the given year (see Table 2.1).

The data originally contained only two multiclass variables: year and region. Related binary variables, such as the use of river water and the use of dam water, were combined to create a single multiclass variable. This was done by first converting each combination that occurred into a unique category (such as river and damn water used, as opposed to the two separate

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.

Variable	**Units**	**Recorded**	**Number of Classes**
Water Used	Mega Litres	5846	
Diesel	Litres	5585	
Biodiesel	Litres	25	
LPG	Litres	958	
Herbicide Spray	Times per year	2026	
Year	Class	6049	10
Disease	Class	6049	2
Region	Class	6049	58
Solar	Kilowatt Hours	622	
Irrigation Type	Class	6049	20
Petrol	Litres	4309	
Slashing	Times per year	2290	
Yield	Tonnes	5935	
Irrigation Energy	Class	6049	16
Area Harvested	Hectares	6049	
Electricity	Kilowatt Hours	1014	
Insecticide Spray	Times per year	1092	
Fertiliser	KGs of Nitrogen	795	
Fungicide Spray	Times per year	2260	
Cover Crop	Class	6049	32
Water Type	Class	6049	39
Grape Revenue	AUD	853	
Operating Costs	AUD	853	

categories prior). These variables were then one-hot-encoded, changing each variable class into a binary value, with one indicating the presence of the class and zero indicating its absence. Further details about classes and their frequency is available in the appendices.

The variable region represented one of the 65 Geographical Indicator Regions (GI Region) used to describe different 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).

2.2. XGBoosted Trees

XGBoosted (eXtreme Gradient Boosting) trees, described in more detail below (and further in the appendix), were created using the XGBoost library (Chen and Guestrin, 2016) in the Python Programming language (G. van Rossum, 1995). XGBoosted trees are a boosted tree ensemble method that can be used to classify classes, or predict continuous response variables. They were chosen for this analysis as the data contained a mixture of class and continuous variables. Moreover, XGBoosted trees are unaffected by multicollinearity, and offer high predictive performance for a wide variety of purposes (Chen and Guestrin, 2016).

XGBoosted models were constructed with operational cost and grape revenue as the predicted variables. The analyses were aimed at uncovering what factors influenced these variables and to what extent. As the purpose of the analysis was to identify relationships between variables and to show how they interact, an XGBoosted tree was created for each of the predictor variables as well. Trees for the predictor variables did not include operational cost or grape revenue as predictors. By creating an XGBoosted tree for each variable it meant that every variable would have a measure of its relative importance to every other variable (see Section 2.3). Together these models were used to measure the interrelationships of the ten most important variables in determining operational cost and grape revenue using variable importance. These measures of relative importance were used to illustrate the highly interrelated nature of variables within vineyards. The interaction between variables was depicted through the use of Sankey and Chord diagrams; with variable importance measures being used to show the strength of connection between the respective predictor variable and the response (see section 2.3).

Due to constraints from the XGBoost library region could only be incorporated as a one-hot-encoded variable when used as a predictor. To better show what variables were related to region overall, another XGBoost tree was created with Region as the predicted value. The difference for this model was that relative variable importance would only be measured for each region specifically, as opposed to a variables overall importance in determining region. Separately profit (the difference between revenue and operational costs) and year was looked at in prior analyses (see appendix) but these results were not included due to low average loss values and model stability.

XGBoosted trees are 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 used to create new decision trees that add to the ensemble. Each new tree created is done so using a loss function that is optimised iteratively to improve upon prior tree's predictive power. The loss function can be any convex function, allowing gradient descent to traverse the loss space until, no improvements can be made via traversal. 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.

2.3. Variable Importance

Due to XGBoost creating a large amount of decision trees, the interpretability of these models is obfuscated by the intricate relationships within complicated ensembles. A measure of variable importance was the technique used to highlight a variables influence within the ensemble. Variable importance can be measured in multiple ways; we used the frequency of a variable appearing as a node within the ensemble as a measure of its importance. This measure was chosen as it connected a variable to the minimisation of its associated objective function. The measure of a variable's importance within this study can then be interpreted as how often a variable was the optimal choice in reducing the loss function of the ensemble. Importantly, multiclass variables being one-hot-encoded (see Section 2.1) are given an importance score for each individual class; for example, each specific region will have its own importance score.

The Sankey and Chord diagrams were constructed using the Holoviews python library (Rudiger et al., 2020). Both Chord and Sankey diagrams illustrated variable importance through the size of the bands between two variables. The number at the end of a connection in a Sankey diagram indicates a variable's importance, or the number of times it appeared within the ensemble. Sankey and Chord diagrams are presented together; with Sankey diagrams showing the connection of a variable to its ten most important predictor variables. Chord diagrams were used alongside a Sankey diagram to show the interconnectedness of the ten most prominent variables within its associated Sankey diagram. Chord diagrams formed circles, with variables being connected through their relative importance. The importance values for the Chord diagrams were taking from the models of those individual variables, with the diagram being simplified to just the ten variables in the associated Sankey diagram, for readability's sake.

2.4. Validation

The predictive accuracy of each tree was assessed through a validation process. For each model the data was split into training data, which constituted 80% of the original data. The remaining 20% was used in testing and validation. Categorical data was stratified to conserve the same proportion of class occurrences between training, testing and validation data. For continuous variables 20% was used as testing data and 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. 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.

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. 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 category variables (Hanley and McNeil, 1982). 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

The prediction of revenue performed similarly to operating cost achieving an R^2 of 0.7716 (with a standard deviation of 0.1525). The value of predictors' relative importance was then calculated through the number of nodes used within the XGBoost. Values for relative importance were then used to construct Sankey and Chord diagrams to compare the contribution of each variable to predicting revenue.

In order of importance, predictors of revenue were fuel use(petrol 307 and diesel 144), yield (285), size (216) and water use (199). Here, the values in the brackets indicate the relative importance of each variable (see C.10). Overall regions contributed to 234 nodes in the ensemble making them collectively the third most important variable. The chord diagram illustrates that vineyard area is also of high relative importance to other variables especially slashing. The overall importance of area to other variables is evident by its larger circumference within the chord diagram (see B in Figure C.10).

3.2. Operating Costs

Comparatively to revenue, operating cost performed better with the XG-Boosted regression ensemble achieving an R^2 of 0.8025 (with a standard

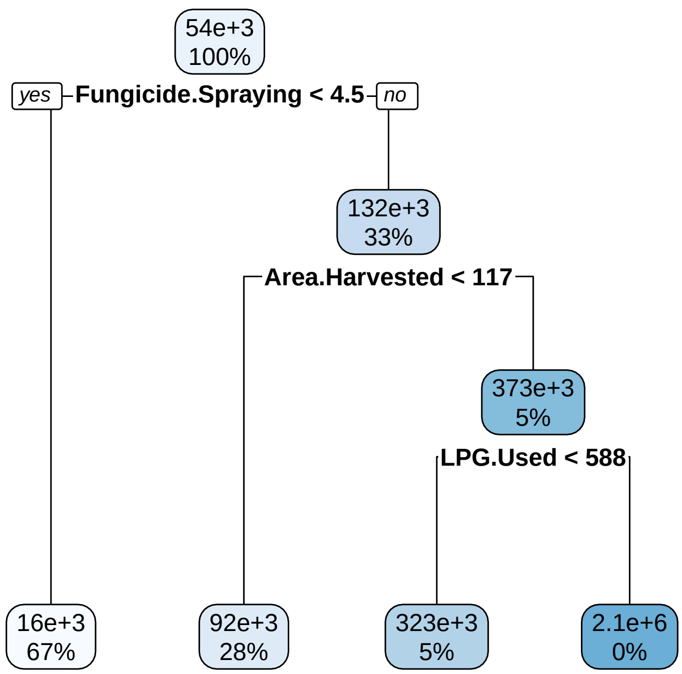

Figure 1: Decision tree predicting revenue. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

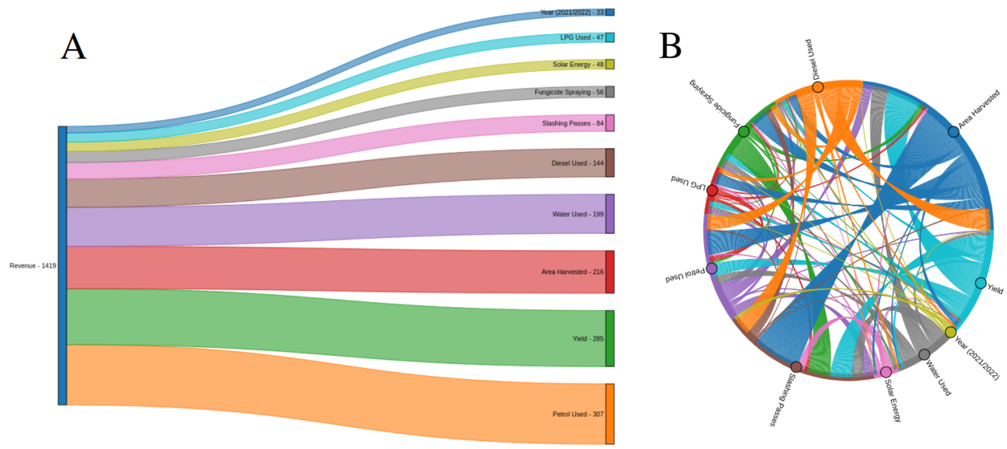

Figure 2: The left-hand side depicts the 10 most important variables in predicting revenue using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

deviation of 0.1033). The relationships to operating cost through variable importance were found to be similar to that of revenue, with fuel, water, area and yield having the largest number relative importance (see figure 4). A surprising difference is that the most important operational consideration for operating cost is the use of fungicide, compared to revenue where slashing is the most important (comparing Figure 6). 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

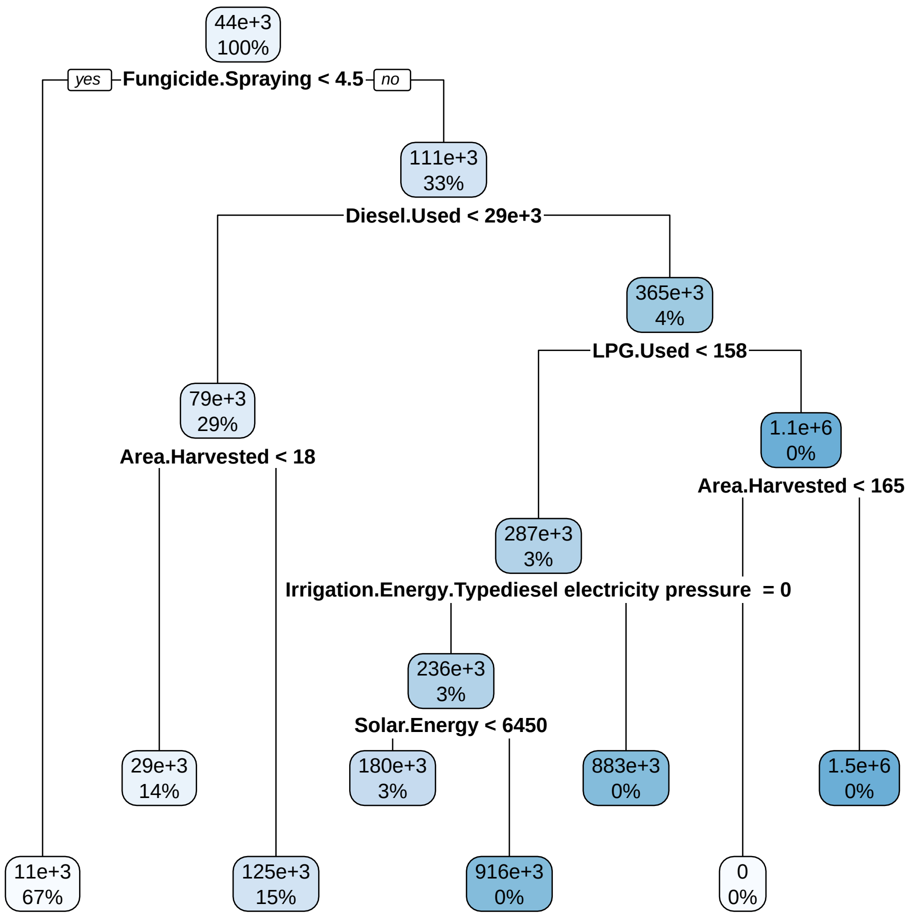

Figure 3: A surrogate model decision tree predicting operating costs. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

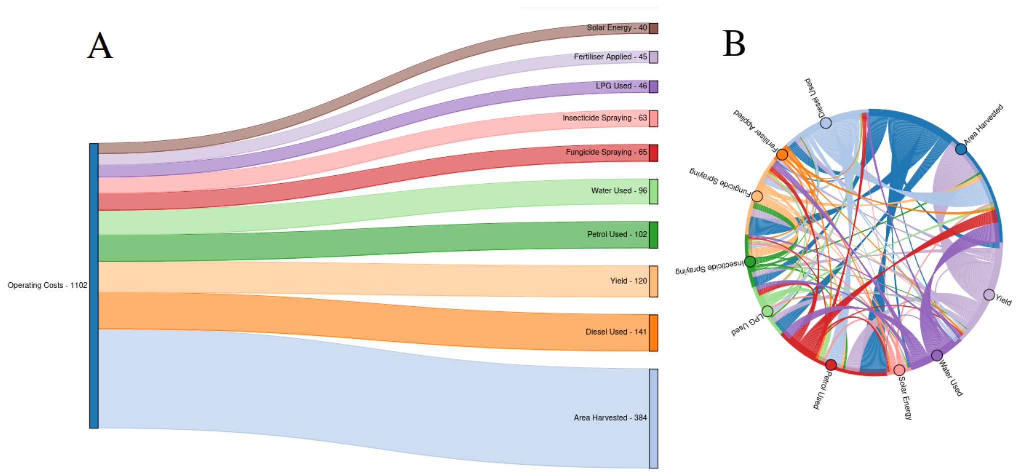

Figure 4: The left-hand side, A, depicts the 10 most important variables in predicting Operating Costs using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The number at the end of each band in the diagram is that variable's importance. The right-hand side, B, depicts the importance of the 10 variables in Sankey diagram relative to one another.

spraying passes were not more prominent in operating costs due to the intrinsic nature as an agricultural expense.

3.3. Region

When considered overall, 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 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.

As also noted above Region was also a key determinant of operating costs. Again Tasmania was the most important, followed by the Adelaide Hills. In contrast to revenue, both climates are considered cool and wet, and warmer drier regions such as the Barossa and Hunter Valley only contributed roughly half the same number of nodes to the ensemble. Based on further analysis of Regions (Figure 6) the inclusion of slashing and fungicide spraying is the likely reason with fungal and weed pressure being greater in cooler wetter regions.

The XGBoost ensemble, did not perform well when predicting operating costs or revenue with 56.82% accuracy (50.58% validation accuracy). The difference in accuracy is in part due to the large number of classes (58 regions). The ensemble had a great emphasis on area, water, fuel and yield as determining factors (see Figure (6).

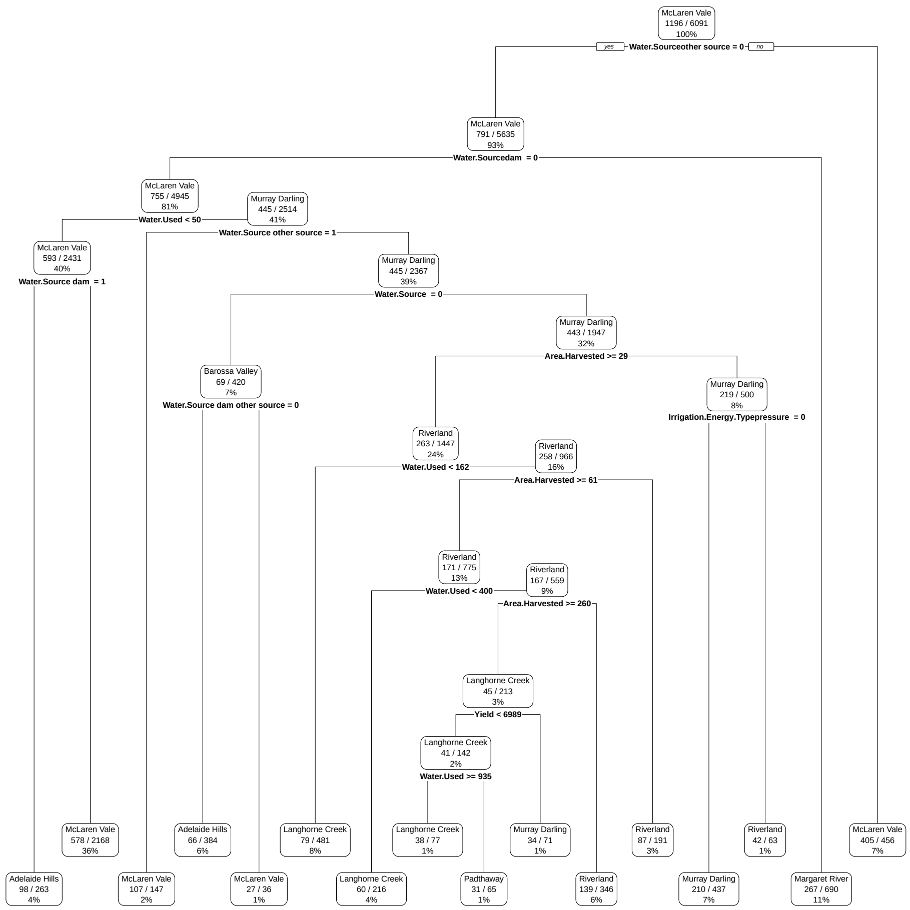

Figure 5: Decision tree predicting Region. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

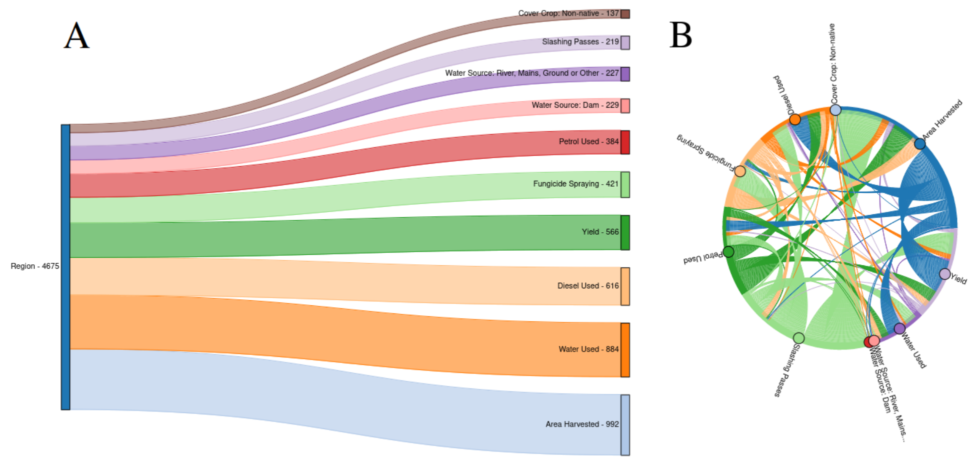

Figure 6: The left-hand side, A, depicts the 10 most important variables in predicting Region using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The number at the end of each band in the diagram is that variable's importance. The right-hand side, B, depicts the importance of the 10 variables in Sankey diagram relative to one another.

Many of the regions had significantly lower reporting rates, resulting in much poorer classification performance. The regions with the most samples performed the best. Bordering regions were routinely grouped together and misclassified as the same region. Two areas that suffered the most from this were the Limestone Coast (cool coastal areas in South Australia) and the warmer inland regions along the Murray Darling.

4. Discussion

This study has explored the relationships between vineyard resource use, operations and geographical properties to revenue and operating costs

geographical regions with respect to a vineyards revenue and operating costs. The analysis was based on a large national study of XXX

Three main findings were identified. First, the most important predicted of revenue were XXX

With YYY showing the relationships among the corresponding set of predictor variables

Second XXX for operating cost

Third for regions...

highlight how decisive regional influences can be determining a vineyard's economic outcomes.

Several physical parameters such as climate, geography 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 association between yield and region is demonstrated by its rank of fourth-highest variable importance when determining region (see Figure 6).

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). In general warmer regions are not associated with higher yields, but if a vineyard in a warmer region is sufficiently irrigated much higher yields can be achieved than in cooler regions (Camps and Ramos, 2012). 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 given more accessible water resources. 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 (and thus quality) (Goodwin I, Jerie P, 1992; MG McCarthy et al., 1986).

In part some winegrowing strategies are restricted simply through access to water resources, being reflected through the region classification tree (see Figure 5). 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 because vineyards in regions without pressurised water will need to use more fuel to pressurise their irrigation systems.

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 node frequency 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 it was attempted to capture the complexity between how variables interacted when determining operational costs (see Figure 4), it is likely yet 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 (via large R^2 values), 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). Reductions in fuel, water and tractor use are obvious methods to reduce operational costs but not necessarily achievable decisions. Without fully capturing more granular activities for example the specific reasons for fuel use, it is difficult to determine what decisions specifically influence the operational costs.

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. 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. Some decisions leading to latent effects such as large scale soil deposition in extreme rain events can be caused by soil compaction due to overworking a vineyard (Capello et al., 2020).

5. Conclusion

This study has provided valuable insights into the multifaceted dynamics governing operational costs and revenue. 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

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.

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.

Botting, D., Dry, P., Iland, P., 1996. Canopy architecture-implications for Shiraz grown in a hot, arid climate .

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G., 2013. API design for machine learning software: Experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122.

Camps, J.O., Ramos, M.C., 2012. Grape harvest and yield responses to inter-annual changes in temperature and precipitation in an area of north-east Spain with a Mediterranean climate. International Journal of Biometeorology 56, 853–64. doi:10.1007/s00484-011-0489-3.

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.

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. Centrum voor Wiskunde en Informatica (CWI),.

Goodwin I, Jerie P, 1992. Regulated deficit irrigation: Concept to practice. Advances in vineyard irrigation. Australian and New Zealand Wine Industry Journal 7.

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.

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.

MG McCarthy, RM Cirami, DG Furkaliev, 1986. The effect of crop load and vegetative growth control on wine quality. .

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.

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. Zenodo. 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.

SWA, S.W.A., 2022. Sustainable Wingrowing Australia. https://sustainablewinegrowing.com.au/case-studies/.

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 .

Appendix A. Continuous variables

The Table below shows the ranges of each of the continuous variables:

	count	mean	std	min	0.25	0.5	0.75
Vineyard Solar	622	22916.89	104808	1	1170.75	5500	14866
Biodiesel	25	6635.932	11768.832104	1	200	500	10000
Fungicide Spray	2260	7.724801	3.279794	1	6	7	9
LPG	958	327.831399	861.538804	1	40	95.835	240
Petrol	4309	825.276809	1556.621119	1	135	306.66	903
Insecticide Spray	1092	1.707189	1.316042	0	1	1	2
Water Used	5846	7301838	558206600	0.0007	13.2655	43	146.8
Fertiliser	795	91149.89	483913.4	1	560	4759.5	45148
Diesel	5585	11677.070183	24380.588742	0.1267	1240	3850	12500
Yield	5935	772.902449	2175.113895	0.03	68	192.3	601.8
Herbicide Spray	2026	2.646199	2.598899	0	2	2	3
Slashing	2290	3.311485	1.826788	1	2	3	4
Electricity	1014	58223.07	177626.3	0.019	2160	9637	36498
Area Harvested	6049	66.52604	133.4525	2.220446E-16	10.13	24.5	66.8
Grape Revenue	875	377972	606286.8	1	76000	172964	38674
Operating Costs	853	314187.1	511522.6	1	57315	140000	32740

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 types

Appendix B.2. cover crop types

Appendix B.3. irrigation type

Appendix B.4. irrigation energy type

Appendix B.5. year

Appendix B.6. giregion

Appendix C. XGBoost

Following Chen and Guestrin (Chen and Guestrin, 2016), XGboosted trees predict 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:

water types
river water
groundwater
surface water dam
recycled water from other source
groundwater and surface water dam
not listed
mains water
river water and groundwater
groundwater and recycled water from other source
other water
river water and surface water dam
groundwater and water applied for frost control
groundwater and mains water
river water and groundwater and surface water dam
recycled water from other source and mains water
groundwater and recycled water from other source and mains water
river water and mains water
surface water dam and mains water
groundwater and other water
river water and groundwater and mains water
groundwater and surface water dam and recycled water from other source
river water and water applied for frost control
groundwater and surface water dam and mains water
surface water dam and recycled water from other source
river water and recycled water from other source
river water and other water
river water and surface water dam and mains water
river water and groundwater and surface water dam and mains water
mains water and other water

Cover crop types
permanent cover crop volunteer sward
permanent cover crop non native
permanent cover crop native
annual cover crop
groundwater and surface water dam
annual cover crop and permanent cover crop volunteer sward
bare soil
permanent cover crop non native and permanent cover crop volunteer sward
annual cover crop and permanent cover crop non native
bare soil and permanent cover crop volunteer sward
bare soil and permanent cover crop non native
annual cover crop and permanent cover crop non native and permanent cover crop volunteer swar
bare soil and annual cover crop
permanent cover crop native and permanent cover crop volunteer sward
bare soil and permanent cover crop native
annual cover crop and permanent cover crop native
permanent cover crop native and permanent cover crop non native
permanent cover crop native and permanent cover crop non native and permanent cover crop volu
annual cover crop and permanent cover crop native and permanent cover crop non native and per
bare soil and annual cover crop and permanent cover crop volunteer sward
bare soil and permanent cover crop non native and permanent cover crop volunteer sward
annual cover crop and permanent cover crop native and permanent cover crop volunteer sward
bare soil and annual cover crop and permanent cover crop native
annual cover crop and permanent cover crop native and permanent cover crop non native
bare soil and annual cover crop and permanent cover crop non native
bare soil and annual cover crop and permanent cover crop native and permanent cover crop non n
bare soil and annual cover crop and permanent cover crop non native and permanent cover crop v
bare soil and annual cover crop and permanent cover crop native and permanent cover crop non n
bare soil and permanent cover crop native and permanent cover crop non native

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
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

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
diesel and solar	4
diesel and pressure and solar and	1

data year id	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

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

, where

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

As predictions are made using additive tree functions, XGboosted trees 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.

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
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

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 (Buitinck et al., 2013), which was utilised via an API to the XGBoost library (Chen and Guestrin, 2016).

Appendix C.2. Year

The classification tree and XGBoosted ensemble 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 perfomed poorly compared to operating cost and revenue 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.

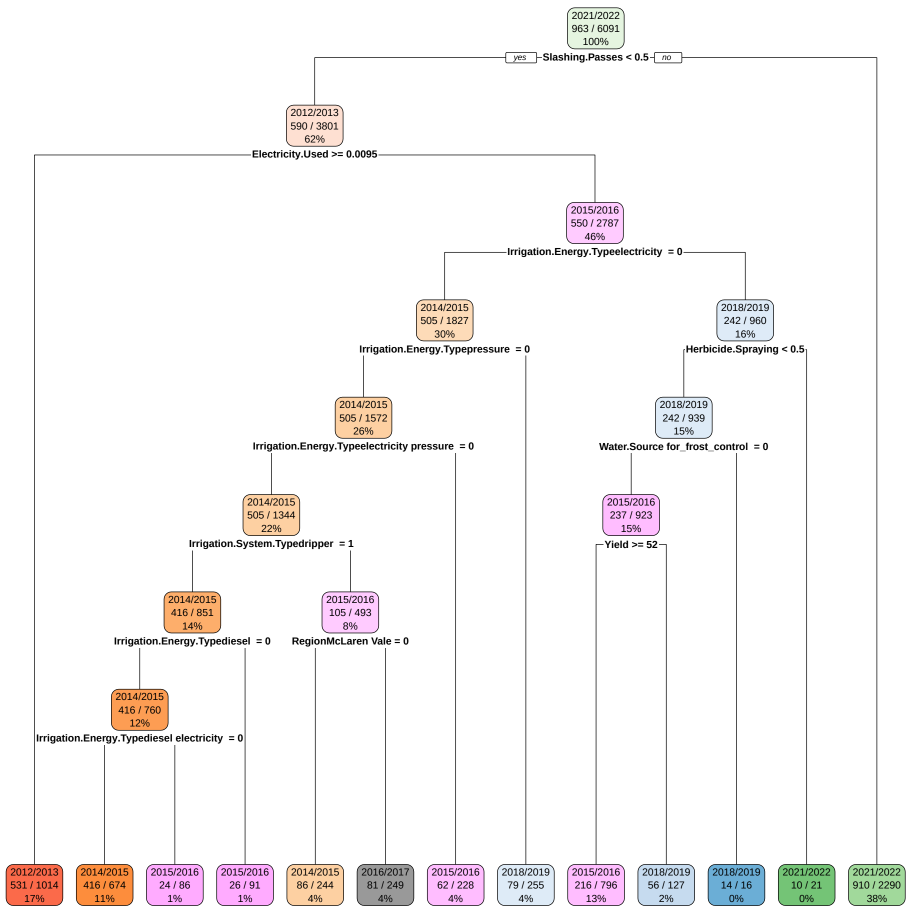

Figure C.7: Decision tree predicting Year. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

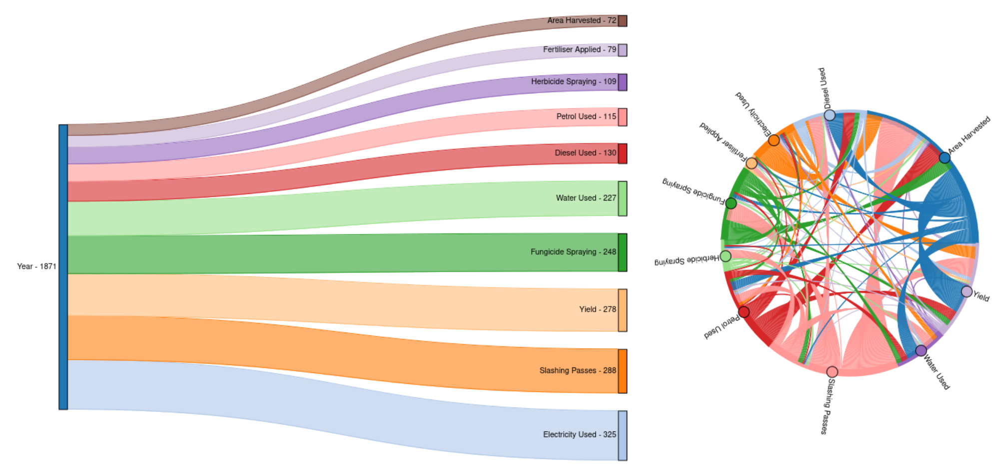

Figure C.8: The left-hand side depicts the 10 most important variables in predicting Year using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

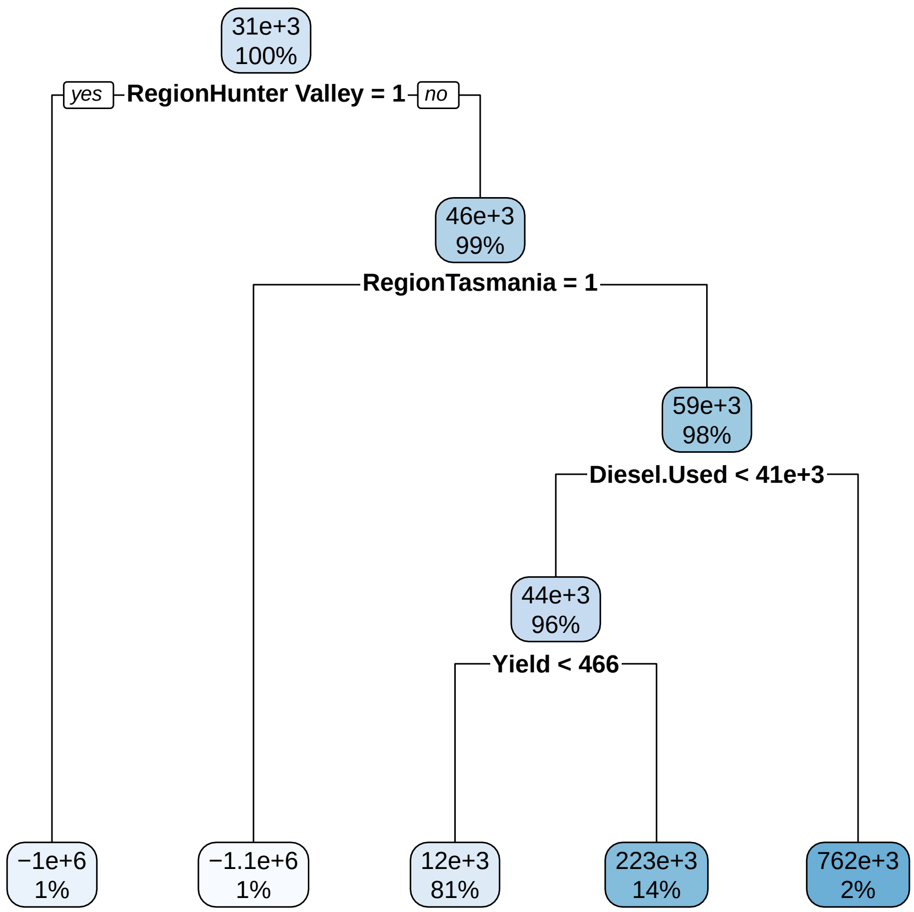

Figure C.9: Decision tree predicting revenue. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

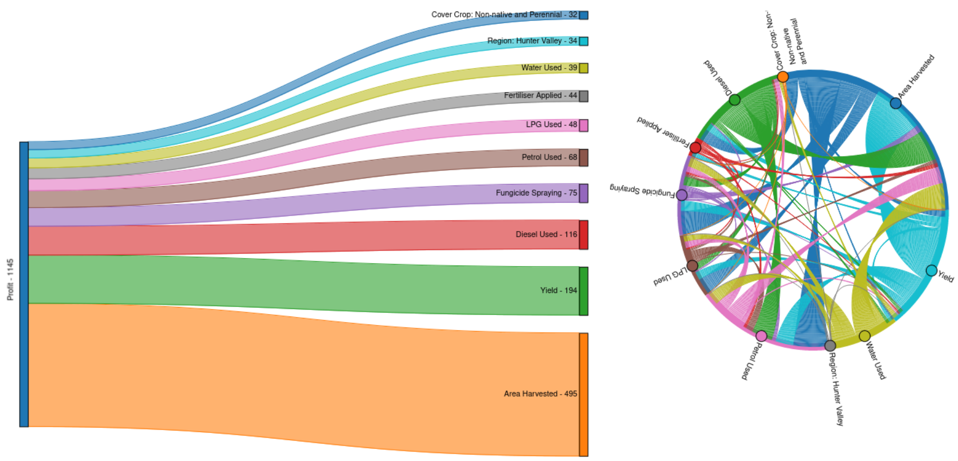

Figure C.10: The left-hand side depicts the 10 most important variables in predicting revenue using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.