Highlights

An analysis of underlying relationships of Australian vineyard's economic outcomes.

Author

- Highlight 1
- Highlight 2
- Highlight 3
- Highlight 4

An analysis of underlying relationships of Australian vineyard's economic outcomes.

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 effected 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 introducing stronger policy and information 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 allowed 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. Answering what the driving factors are behind vineyard economic outcomes, and linking these outcomes to predictor importance. This is done through analysing a new comprehensive nationwide data set using XGBoosted models. We further compare the relationships between different resources to address the extensive collinearity found within the data (Chen and Guestrin, 2016). XGBoosted models were used because they are able to overcome multicollinearity as well as highlight the level of importance that predictor variables have on response variables; with importance being able to be statistically defined through multiple methods.

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 facilitate grape-growers and winemakers in demonstrating and improving their sustainability (SWA, 2022). Data recorded by SWA is entered manually by winegrowers using a web based interface tool. A total of 6091 observations were collected from 2012/2013 to 2021/2022 financial years. 23 variables were used for each observation reflecting a vineyards state for the given year (see Table 2.1).

The data originally contained only two multiclass variables: year and region. Variables that measured the same metric from different sources (such as

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	6091	10
Disease	Class	6091	2
Region	Class	6091	58
Solar	Kilowatt Hours	622	
Irrigation Type	Class	6091	20
Petrol	Litres	4309	
Slashing	Times per year	2290	
Yield	Tonnes	5935	
Irrigation Energy	Class	6091	16
Area Harvested	Hectares	6091	
Electricity	Kilowatt Hours	1015	
Insecticide Spray	Times per year	1092	
Fertiliser	Kilograms of Nitrogen	795	
Fungicide Spray	Times per year	2260	
Cover Crop	Class	6091	32
Water Type	Class	6091	39
Grape Revenue	AUD	853	
Operating Costs	AUD	853	

water collected from rivers versus water from dams) were converted into multiclass variables representing the source through one-hot-encoding. Changing each variable class into a binary value, with one indicating the presence of the class and zero indicating its absence. Occurrences of multiple sources were defined as their own separate classes. Where a class variable had a recorded amount the total amount used from these variables was retained as a separate variable; for example water used (in Mega Litres) was also included alongside water source.

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 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. And, XGBoosted trees are unaffected by multicollinearity, as well as offer high predictive performance for a wide variety of purposes (Chen and Guestrin, 2016). An XGBoosted tree was created for each variable to show how they interacted. Each tree included all but the economic variables (operating cost and revenue from grape sales), which were only included within their own trees as response variables. Separately profit (the difference between revenue and operational costs) was looked at in prior analyses (see appendix) but the results were not included due to low average loss values and model stability. This meant that every variable would have a measure of its importance to other variables (see Section 2.4), which was used to show the highly interrelated nature of variables within vineyards. The complicated interaction between bariables was illustrated using Sankey and Chord diagrams; with variable importance measures being used to show the strength of connection between any two variables (see section 2.4).

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:

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

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

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

, where

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

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.

2.3. Loss functions

The functions included as parameters in equation 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:

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

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 (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).

2.4. 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 are given an importance score for each individual class; for example, each specific region will have its own importance score.

Creating XGBoosted trees for each variable allowed the use of importance to show how strongly variables were associated with each other. The importance of variables to one another was illustrated through the use of Sankey and Chord diagrams. These 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 indicated a variable's importance, or the number of times it appeared within the ensemble. Sankey and Chord diagrams were presented together; with Sankey diagrams showing the connection of a variable to its 10 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.

2.5. Validation

As there were multiple different loss functions, multiple different forms of validation were used. In each case the data was split into training data, which constituted 80% of the original data. The remaining 20% was used in testing and validation. Data was stratified when splitting the data into these subsets to conserve the same proportion of class occurrences between training, testing and validation data. For continuous variables 20% was used as testing data, minimising the root-mean-square function. The final model was validated using repeated k-fold cross validation for 10 folds, repeated 10 times. R^2 scores were used to determine the best regression models during validation. For binary and multiclass variables, data was split into 80%

training, 10% testing and 10% validation data. For class 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).

2.6. Surrogate Models

The creation of more interpretable models such as linear regression in parallel to AI systems has been used to explain variable's relationships within black box algorithms (Molnar, 2022). As XGBoost create an ensemble of decision trees, here we use classification and regression trees to gain insight into intricacies of the ensembles derived through XGBoost. Decision Trees were created for operating costs, revenue and region. These models describe the partitions that are useful in predicting these variables; giving insight into the trees that make up the ensembles created by XGBoost. These trees were created using the rparts and caret packages (Kuhn, 2008; Terry Therneau and Beth Atkinson, 2022) in the R statistical programming language (R Core Team, 2021).

Decision trees were validated using K-fold cross validation. Each model was validated using 10 folds, utilising a random selection of different samples ten separate times to validate each of the decision trees. The same measure of accuracy as the XGBoosted trees was used for comparison.

3. Results

3.1. Revenue

We investigated the link between revenue to other variables in the SWA data by predicting it, and then linking each variable to revenue through variable importance. The prediction of revenue performed similarly to operating cost achieving an R^2 of 0.7716 (with a standard deviation of 0.1525). A regression tree was used as a surrogate model to present an example of the typical type of decision tree present within the XGBoost Ensemble, however the surrogate model only achieved an R^2 of 0.0961 (with a standard deviation of 0.0181) and the XGBoosted ensemble.

The important variables when attempting to determine revenue were size, yield, fuel and water (see .10). Due to regions being recorded separately for importance none appeared as the most important variables, overall regions contributed to 234 nodes in the ensemble making them collectively the third most important variable. Although performing poorly, the surrogate model highlights the importance of size in determining revenue. Area also appearing as a variable of higher importance is show to be highly interrelated with other variables. The relation to area is likely to primarily be the effect of economies of scale, shown through its strong relations to other variables in figure .10. Area harvested is likely also an indicator of other variables such as slashing passes its strongest connection presented.

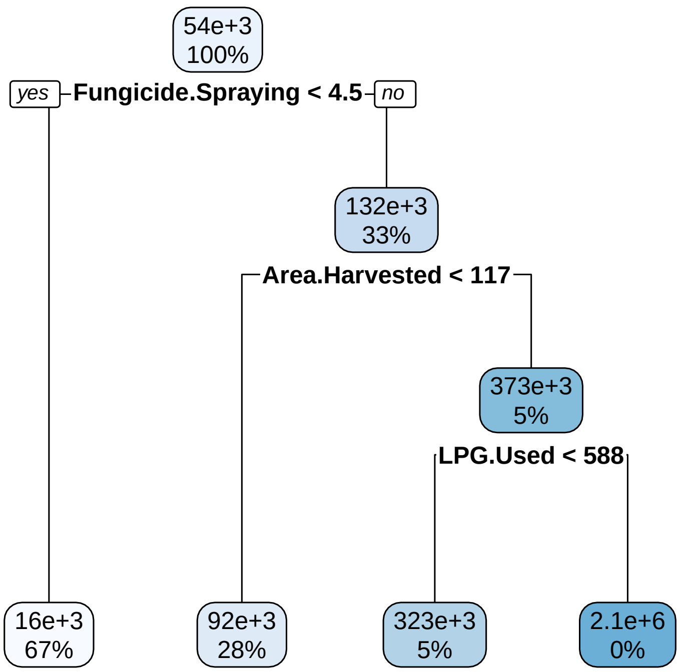

Figure 1: Decision tree predicting revenue. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

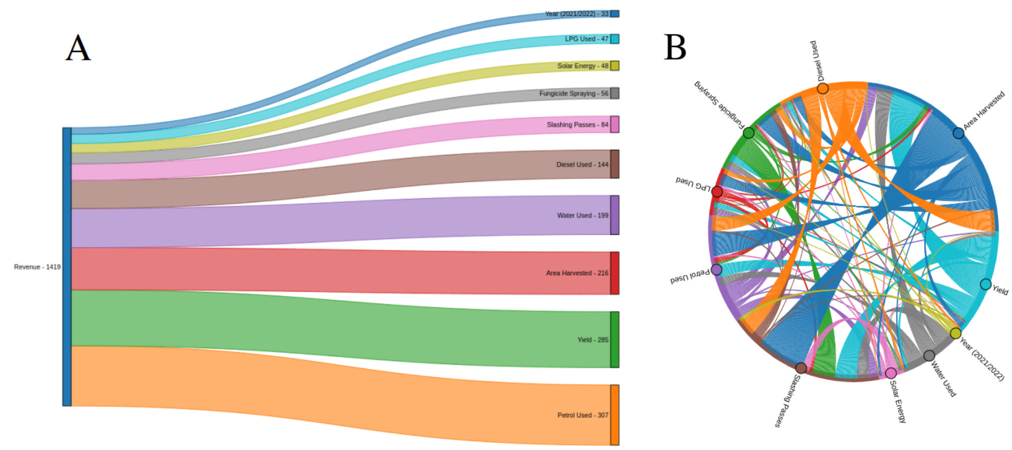

Figure 2: The left-hand side depicts the 10 most important variables in predicting revenue using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

3.2. Operating Costs

The relationships to operating cost through variable importance were found to be similar to that of revenue. With fuel, water, area and yield occurring the most (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 and 4). The variables that feed into these decisions are also very differing with diesel being the most informative to slashing and area being the most informative to the need for fungicide.

Again, region played a determining factor overall but not as much individually with region contributing to 334 nodes within the ensemble making it the most important variable when considering all regions together. It

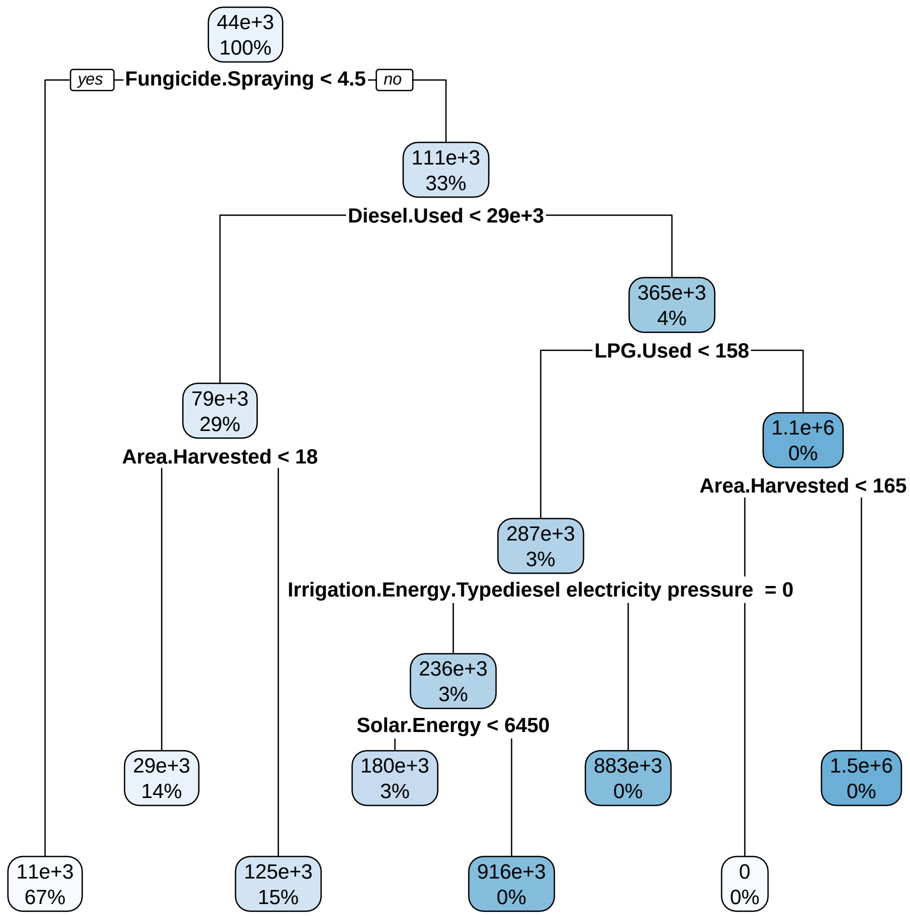

Figure 3: A surrogate model decision tree predicting operating costs. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

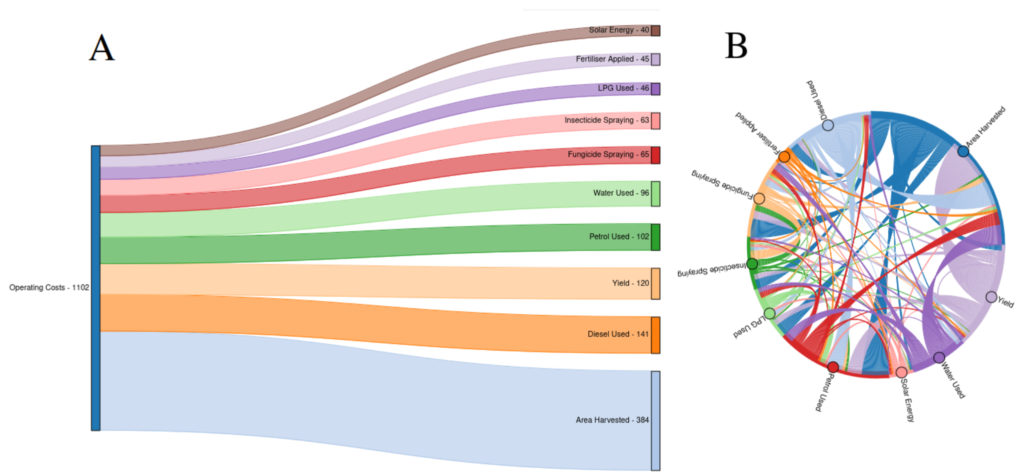

Figure 4: The left-hand side, A, depicts the 10 most important variables in predicting Operating Costs using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The number at the end of each band in the diagram is that variable's importance. The right-hand side, B, depicts the importance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

was surprising that electricity, slashing and spraying passes were not more prominent in operating costs due to the intrinsic nature as an agricultural expense.

Comparatively to revenue, operating cost performed better The XG-Boosted regression ensemble achieved an R^2 of 0.8025 (with a standard deviation of 0.1033). Again the surrogate model did not perform well achieving an R^2 of 0.0931 (with a standard deviation of 0.0197) but showed similarly to revenue an importance placed on fungicide spraying and size (see figure 3).

3.3. Region

When considered overall, Region was a highly informative variable through measure of importance to both operating cost and revenue. Notably the overall importance of region to revenue was 234 (making it the third most important variable when considerng all regions together). 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 being warm and dry climate focussing on Shiraz grapes and Tasmania being a cool wet climate that grows Pinot.

When considering all regions together, it had the most node contributions to determining operating costs with an importance of 334. Of all the regions, 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

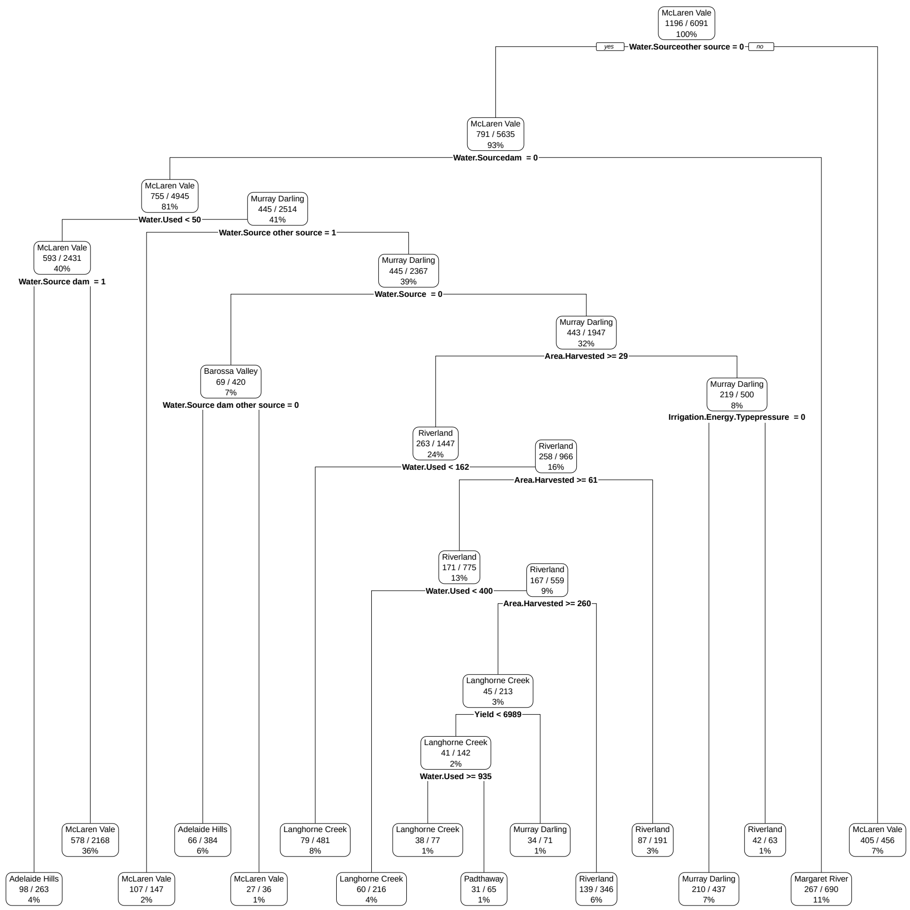

Figure 5: Decision tree predicting Region. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

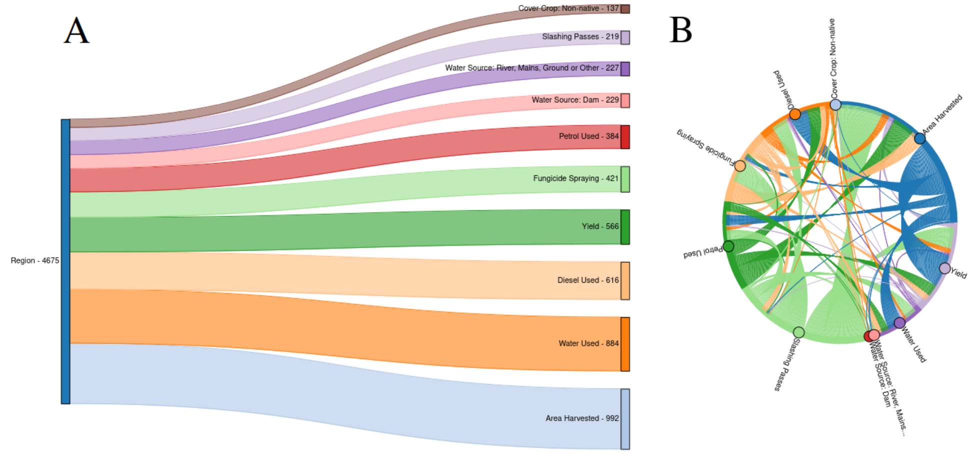

Figure 6: The left-hand side, A, depicts the 10 most important variables in predicting Region using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The number at the end of each band in the diagram is that variable's importance. The right-hand side, B, depicts the importance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

half the same number of nodes to the ensemble. When looking at 6 the inclusion of slashing and fungicide spraying is the likely reason; with fungal and weed pressure being greater in cooler wetter regions.

Both diesel and petrol were of more relative importance in operating costs than water was compared with region. similar to those used to classify region (see Figure 6, except water used. The surrogate model relied heavily on the use of water source to classify regions, which is reflective of regional access to resources being a deciding factor in vineyard management (see Figure 5). A major difference between region and revenue was the importance given to water use, with water use being a relatively more important variable in predicting region than revenue (considering its rank in importance to other variables).

The surrogate model for region performed better than other surrogate models with 32.34% (3.67% standard deviation). The prominence of types and use of water resources was in classifying region is reflective of difference of availability of water resources is when comparing different regions (see Figure 5). The XGBoost ensemble, did not perform as well as 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 (being 58). The ensemble did not differ greately from the surrogate model, with a continuing emphasis on Area, water, fuel and yield as determining factors (see Figure (6).

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, specifically with the classification tree were the Limestone Coast (cool coastal areas in South Australia) and the warmer inland regions along the Murray Darling. The classification tree likely had more difficulty discerning vineyards closer to the river using only water sources due to the greater access to river water in these areas.

4. Discussion

The explored relationships between vineyard resource use, operations and geographical properties to revenue and operating costs 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. Where, the need to pressurise irrigation systems from river water or to generate power would require larger amounts of diesel and petrol.

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 more circumstantial in determining operational costs. 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 a vineyards Geographical Indicator region.

Decisions made on the ground have far-reaching effects and are difficult to completely capture. A higher 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. With factors such as erosion and soil health being difficult to capture but also influenced by tractor use (Capello et al., 2019, 2020). Although, performing well in R^2, the ability to predict operational costs is limited by the variables incorporated. 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 such as the specifics of what fuel was used for, it is hard to determine what decisions specifically influence the operational costs. Electricity in particular is used predominantly for irrigation. Size is also a further consideration where slashing and spraying are measured in discrete tractor passes and show a surprising connection to the overall size of a vineyard, despite not being scaled to any measure of size. This would mean that, although measured as the same increment, a slashing or spraying pass in a larger vineyard would consume more fuel and wages than in a smaller vineyard.

The reasoning for any particular decision can be widely varying. A more granular definition of region may help to better discern the differences in practices, and the reason for employing them. 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 decisions made and chance. 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 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.

Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software, Articles 28, 1–26. doi:10.18637/jss.v028.i05.

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. .

Molnar, C., 2022. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2 ed.

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.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

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/.

Terry Therneau, Beth Atkinson, 2022. Rpart: Recursive Partitioning and Regression Trees.

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 .1. 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 .2. 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

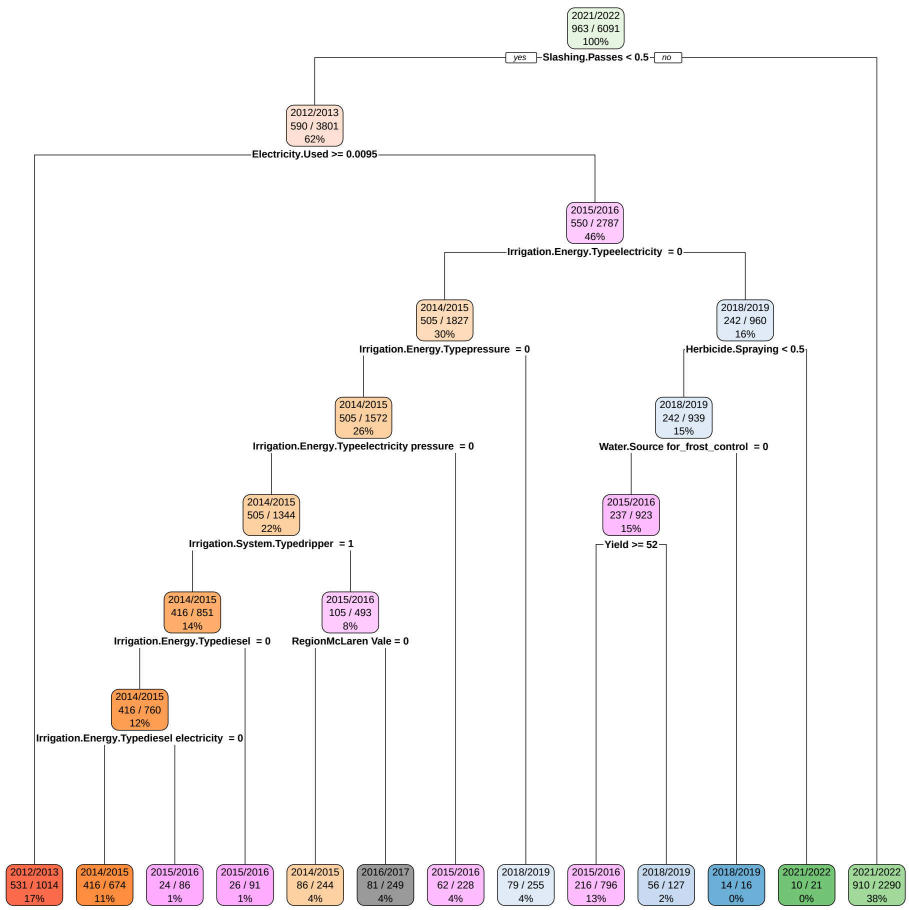

Figure .7: Decision tree predicting Year. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

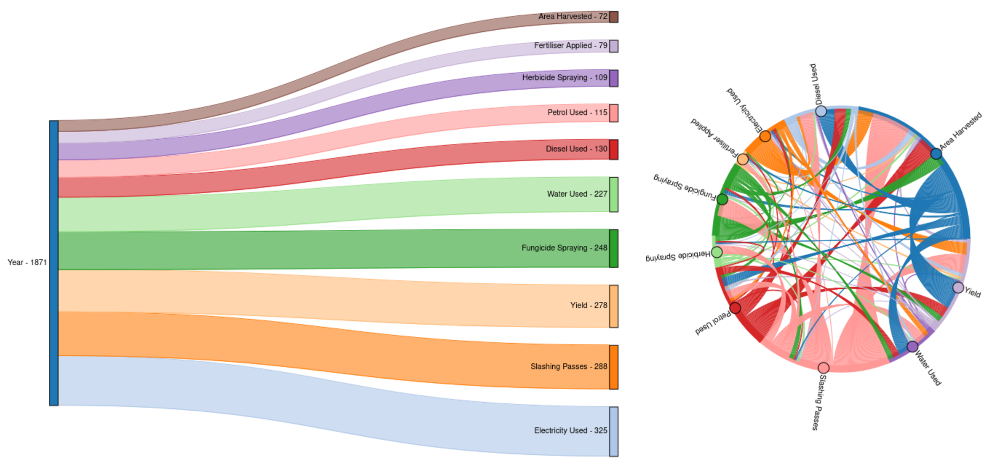

Figure .8: The left-hand side depicts the 10 most important variables in predicting Year using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

the large standard deviation being indicative of how unstable the models created were.

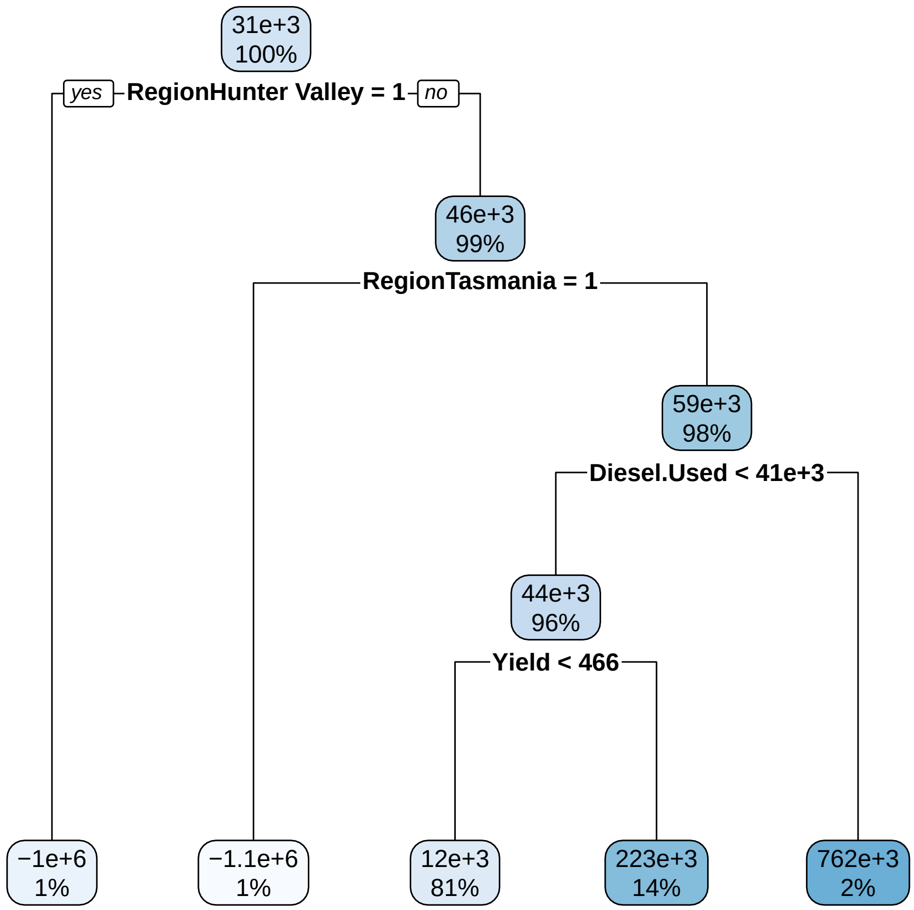

Figure .9: Decision tree predicting revenue. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

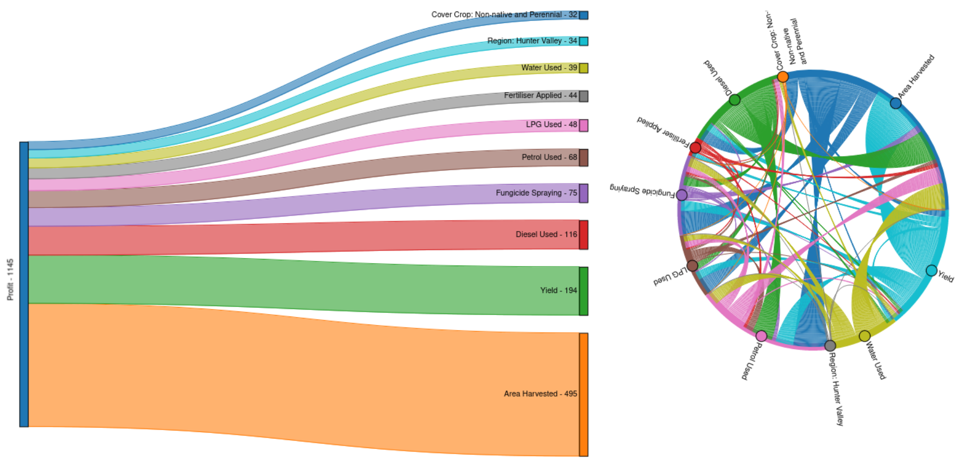

Figure .10: The left-hand side depicts the 10 most important variables in predicting revenue using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.