Syndicate 8

Tony Trinh (1099433)  
Alvin Lee (1169514)  
Varun Sharma (959620)  
 Jonno Lindsay (1026965

Predictive analytics

Syndicate Task #5 – Predicting Wine Quality

Table of Contents

[1. Purpose 2](#_Toc55074669)

[2. Results & Analysis 2](#_Toc55074670)

[2.1 Logistic Regression 2](#_Toc55074671)

[2.2 Boosting 3](#_Toc55074672)

[2.3 Neural Network 3](#_Toc55074673)

[2.4 KNN 4](#_Toc55074674)

[2.5 Random Forest 4](#_Toc55074675)

[3. Comparisons between methods and models 5](#_Toc55074676)

[3.1 Differences between Methods 5](#_Toc55074677)

[3.2 Non-Classification Methods Advantages/Disadvantages 5](#_Toc55074678)

[3.3 Classification Methods Advantages/Disadvantages 5](#_Toc55074679)

[4. Conclusion & Recommendation 6](#_Toc55074680)

[Appendix 7](#_Toc55074681)

[Appendix 1 – Wine loss maker function (modified for classification) 7](#_Toc55074682)

[Appendix 2 – Logistic Regression User Loss versus Cutoff 7](#_Toc55074683)

[Appendix 3 – Boosting In-Bag Risk Reduction 7](#_Toc55074684)

[Appendix 4 – Boosting User Loss versus Cutoff 8](#_Toc55074685)

[Appendix 5 – NN User Loss versus Cutoff 8](#_Toc55074686)

[Appendix 6 – KNN User Loss versus Cutoff 9](#_Toc55074687)

[Appendix 7 – Random Forest Loss User Loss versus Cutoff 10](#_Toc55074688)

[Appendix 8 – Olden Variable Importance – Regression vs Classification 10](#_Toc55074689)

# 1. Purpose

The purpose of this report is to explore whether classifying wine quality scores to good versus bad is a better predictive approach, from statistical and business viewpoint.

The report elaborates how well these classification methods perform and the tradeoffs that can be expected when using them. It will also discuss if running classification yields a better predictive outcome than running models on numeric quality scores. It also aims to determine the advantages and disadvantages a winemaker is expected to face when deciding which group of methods to use and the implications on their business.

# 2. Results & Analysis

The wine makers function has been modified to penalize false negatives quite heavily as a false negative means a large loss in potential revenue. False positives are penalized to a lesser extent, mainly to account for marketing costs. Each false negative is estimated to cost the business $15 and each false positive is expected to cost $2. The function has been expressed per 100 bottles of wine to give a better sense of the associated costs. Further reasoning can be found in Appendix 1.



The table above provides the predictive performance of the models run at the cutoff point which minimizes the predicted loss per 100 bottles.

## 2.1 Logistic Regression



The logistic regression model yielded similar findings to models which had been run in previous syndicates. It highlighted that the most important variables in determining wine quality were ‘Sulphates’, ‘Alc’, ‘TSD’, ‘Ch’, ‘VA’, ‘RS’ and to a lesser extent ‘Density’.

Higher levels of ‘Alc’, ‘Sulphates’ and ‘RS’ all had the effect of increasing the probability of a predicted wine being a good wine, whilst ‘TSD’, ‘Ch’, ‘VA’ and ‘Density’ had the effect of reducing the probability a predicted wine would be a good wine.

Overall, the model performs decently with an AUC of 0.9340. The model at a cut off of 0.21 has a relatively poor overall accuracy of 0.8625, which was the second lowest compared to all other models, performing even worse than the naïve accuracy. However, it performed the third best in the user loss function, with estimated losses of $39.69 per 100 bottles of wines. This is due to the model having a low precision of 0.3881, indicating a tendency to predict more wines as being ‘good’ quality even if they’re not, causing overall accuracy to be poorer but causing the user loss function to predict relatively well.

Appendix 2 demonstrates the varying losses as cut off is changed. The logistic regression model is relatively flexible in the sense that if the cut off is adjusted to a lower or higher value, it is fairly forgiving and the user losses will remain around $40 - $50 per 100 bottles of wine, which should give confidence to the winemakers that it might be a reliable model to use if consistency of losses is what they seek.

## 2.2 Boosting



Although the boosting model uses logistic regression as a base model, it yielded some similarities and some differences. ‘Alc’ and ‘Sulphates’ were the variables which were the most important identical to the logistic regression, however in the boosting model, ‘Alc’ was deemed to be the most important. This was followed by ‘VA’, ’TSD’, ’Ch’, ’pH’ and ‘CA’.

‘pH’ was included in the boosting model, whilst it dropped off ‘RS’ and ‘Density’. This could be partially attributed to the learning rate selected.

Directionally, the boosting model indicated the same results as the logistic regression. Increases in ‘Alc’, ‘Sulphates’ and ‘CA’ increased the probability of a predicted wine being a good wine and an increase in ‘VA’, ’Ch’, ’TSD’ and ‘pH’ decreased the probability a wine would be considered a good wine.

In terms of impact, Appendix 3 demonstrates that ‘Alc’ is by far the most impactful variable followed by ‘VA’, reducing the loss function by approximately 0.62% and 0.23%. This is very similar to the results of the boosting of a linear regression which identified ‘Alc’ as reducing loss function by 0.75% and ‘VA’ reducing it by 0.26%.

At a cut off of 0.27, the boosting model performed averagely across all metrics. It performed slightly better than the logistic regression model in terms of overall accuracy (0.8875 vs 0.8625) but slightly worse in the AUC (0.9265 vs 0.9340) and slightly higher predicted losses per 100 bottles of wine ($42.81 vs $39.69). However, this could be attributed to the boosting model not having the learning rate (nu) being fully optimized.

Appendix 4 highlights the change in user loss costs as the cut off changes. A cut off of 0.27, provides peak performance for this model. In this instance, it is better for the business to have a cut off around 0.27 or slightly lower as the losses do not increase as steeply compared to going beyond 0.27 where the user losses start to increasing quite rapidly.

## 2.3 Neural Network

The Neural Network (NN) for classification reveals some insights using Olden’s variable importance when compared to NN for regression. For instance, ‘FSD’ has changed from a slightly positive variable in regression to extremely negative in classification (-1002.6431), taking the place of ‘Ch’. On the other hand, Sulphates has gone from a minor negative factor in regression to the most positive factor (1397.4927) in classification. Other explanatory variables such as ‘VA’, ‘CA’, ‘RS’, ‘FA’ show similar patterns and relative importance magnitudes across regression & classification. Further details are available in Appendix 5.

At the optimal cutoff of 0.09, NN performs well in terms of overall accuracy (0.9125) and precision (0.5128) compared to other models. However, its AUC (0.8938) is sub-optimal, and the winemaker loss ($54.06) is far from being the top model which emphasises the fact that there is a difference between statistical loss and business loss.

It’s also worth noting that the NN’s winemaker loss comprises of several discrete regions as the cut off value changes. As shown in Appendix 6, the loss value increases marginally from $50 to $70 as the cut off moves from 0.01 to 0.65. However, when that threshold is passed, the loss skyrockets to circa $100. As such, it is critical to use a cut off below 0.65 if the purpose is to minimise winemaker’s loss.

## 2.4 KNN

The KNN model yielded a below average winemaker’s loss of $41.56 per 100 bottles with a decent AUC of .9217 at the cutoff of 0.2. However, the overall accuracy of 0.8531 and precision of 0.3714 were the lowest compared with other models. The model (Appendix 7) is not flexible with respect to the cut off and the loss escalates quickly when the cut off is either increased or decreased from the optimum value of 0.2. The model over predicts the wine quality due to a low precision score but this is not a major concern as over prediction or false positives are preferred to underprediction and false negatives. At the same time, due to an average sensitivity score of 0.8966, the model tends to underpredict or provide false negatives for some good quality wines.

## 2.5 Random Forest

The random forest model’s variable importance is similar to previously run models, where ‘Alc’ and ‘Sulphates’ are the most important, ranking 1st and 2nd respectively. ‘Density’ and ‘VA’ are listed as the next most important variables.

Compared to the non-classification Random Forest model run in session 4, the importance of ‘Alc’ and ‘Sulphates’ are consistent in that they both rank 1st and 2nd, respectively. The only difference is that in the Random Forest classification model, ‘Density’ ranked above ‘VA’ and ‘TSD’. However, this is due to the criteria selected. When the model was run with the same tree, mtry and node selection criteria as per the model in session 4 (300 trees, 9 mtrys and 5 modes), both models listed ‘VA’ and ‘TSD’ being the top important factors.

The model has an overall accuracy score of 0.8688, which is below the naïve model. From an AUC score perspective, the model performs the strongest with a score of 0.9698 and it also has the second-best user loss function, estimating a loss of $25.65 per 100 bottles of wine. However, similarly to the Logistic Regression, the precision score for this model is one of the worst scores (0.4085), due to th of having a lot of false positives.

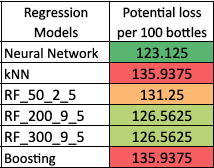
As depicted in Appendix 8, the cut off selected plays a pivotal role in minimising the winemakers’ potential loss. Subject to which cut off point is selected, the loss ranges from $26.25 to $136.88 per 100 bottles of wine. Within the cut off 0.1 to 0.3, the loss holds a similar value ($25-$40), however either side of this, the loss gradually increases towards $100 and over. Therefore, when, utilising the Random Forest method, the cut off which performs the best for the winemaker is 0.17 which results in a loss of $26.25 per 100 bottles.

# 3. Comparisons between methods and models

## 3.1 Differences between Methods

Comparing models between classification and non-classification methods yielded similar findings to the extent that important variables were able to be identified and directional impacts of variables on quality of wines could be assessed.

The below table is derived from modifying syndicate 4’s loss function to something similar to that of the classification loss function. It demonstrates that when using the new loss function per 100 bottles of wine, non-classification (regression) methods score in the range of $120 - $135 of loss, much worse than that of the classification methods.



## 3.2 Non-Classification Methods Advantages/Disadvantages

Non-classification methods have an advantage in that a wine quality score can be produced and distances between quality of wines can be expressed, meaning more granular analysis and pricing can occur. A wine with a predicted score of 7 versus 8 versus 9 could all be treated differently and marketed appropriately. Loss functions can also be tailored to a more granular extent depending on business requirements.

There is also an inherent advantage of being able to explain causation and how different variables impact the final score. For example, the effects of adding x amount of alcohol potentially has on the quality score as opposed to a probability. However, some of the machine learning methods will not be able to explain causation, regardless of whether the method is classification/non-classification.

Building on our two-step approach discussed in the last syndicate exercise, one of the issues with non-classification methods is the tendency to focus on a quality score. It may not be that important for a winemaker to determine if a score is 7.5 versus a score of 9, considering high quality wines should be assessed independently using expert advice and qualitative information. It may be better to simplify the predictive approach and simply separate the wine by a good (≥ 7) and bad (< 7) classification, then verify with independent assessment.

## 3.3 Classification Methods Advantages/Disadvantages

Classification methods provide a probability of a wine being good or bad which can be viewed as an inherent advantage. It may be more important for a business to determine how correct their prediction is as opposed as to a fixation on a quality score. It also shines in the sense that crude groupings of yes/no may be good enough for a business to act on and that distances between quality of wine are irrelevant.

However, an issue with classification methods is their inability to provide granular scores. Some winemakers may see it as crucial in being able to determine distances between wine so prices can reflect those distances and subsequent marketing and pricing strategies can be implemented. Loss functions created for classification methods are also less sophisticated, with less ability to tailor the loss function due to the lack of distances.

Another major issue with classification methods is the selection of cut off points. Given every model produces different probabilities for the same wine, cut off points will differ depending on the loss function and it can be hard to explain why a wine made a cut off in one model but not in another.

# 4. Conclusion & Recommendation

Applying the new loss function on non-classification methods has reveal that the preferred methods for a winemaker would be the classification methods which creates fewer potential loss per 100 bottles. That said, both families of methods can be used to segment good wines and bad wines.

Where the main difference occurs is the ability for non-classification methods to provide a distance between good wines and distances between bad wines. However, consistent with recommendations provided in the previous syndicate, being able to produce distance scores are not so relevant for good wines as it has been recommended to assess these wines in conjunction with a subjective assessment based on experience and other qualitative data.

The classification models provide better actionable information to winemakers for improving and optimizing revenue. This is mainly due to the following reasons:

1. The data set available contains limited data on high quality wines of scores 8 and above.
2. Non-classification models are not very good at extrapolating the data of the wine quality score of 8 and above and hence tend to underpredict score for high quality wines.
3. Wine quality scores are very subjective to the wine tasters and conditions on the day of the wine tasting; hence numerical prediction of wine quality cannot be precise.

Classification models can help winemakers overcome the shortcoming in the data and subjectivity with respect to the wine tasters better. Based on predictions from classification models, it is recommended for winemakers to use Random Forest model as it has a lowest loss and tends not to underpredict.

# Appendix

# Appendix 1 – Wine loss maker function (modified for classification)

The loss function for classification models was developed to calculate the winemaker's loss per bottle of wine with the below assumed pricing model based on the wine quality score.

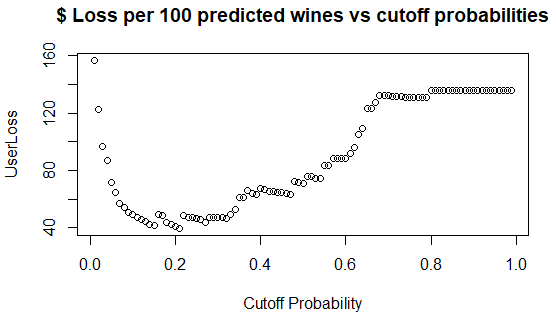
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Quality Score** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Retail Price** | $ 5-7 | $ 5-7 | $ 7-8 | $ 7-8 | $ 8-10 | $ 8-10 | $20 | $30 | $40 | $50 |

It was further assumed, that a wine scoring good wine quality score of 7 will have the following probability to score beyond up to the quality score of 10.

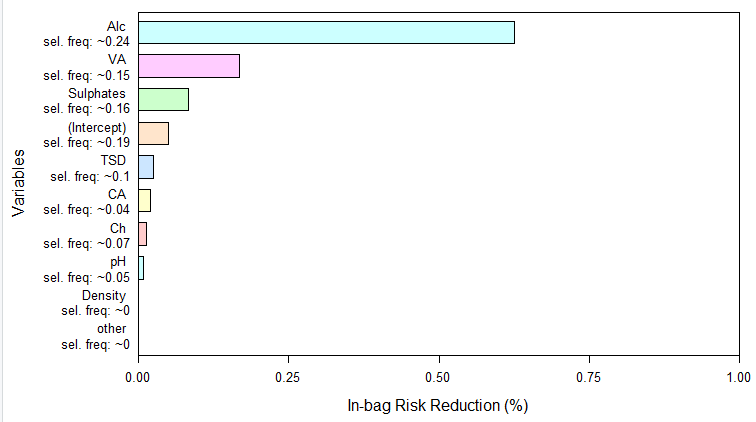
|  |  |  |  |
| --- | --- | --- | --- |
| **Quality Score** | 8 | 9 | 10 |
| **Probability** | **40%** | **15%** | **5%** |

Based on the above assumptions it was concluded that a winemaker can loose on average $15 dollars for underscoring a wine. The loss for overestimating the quality score is estimated $2 per bottle, based on the assumption that winemakers spend 20% or $2 per bottle to market high quality wines.

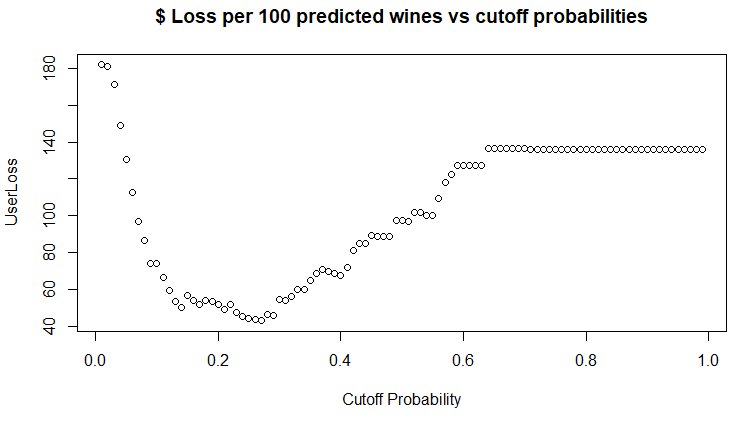
# Appendix 2 – Logistic Regression User Loss versus Cutoff



# Appendix 3 – Boosting In-Bag Risk Reduction

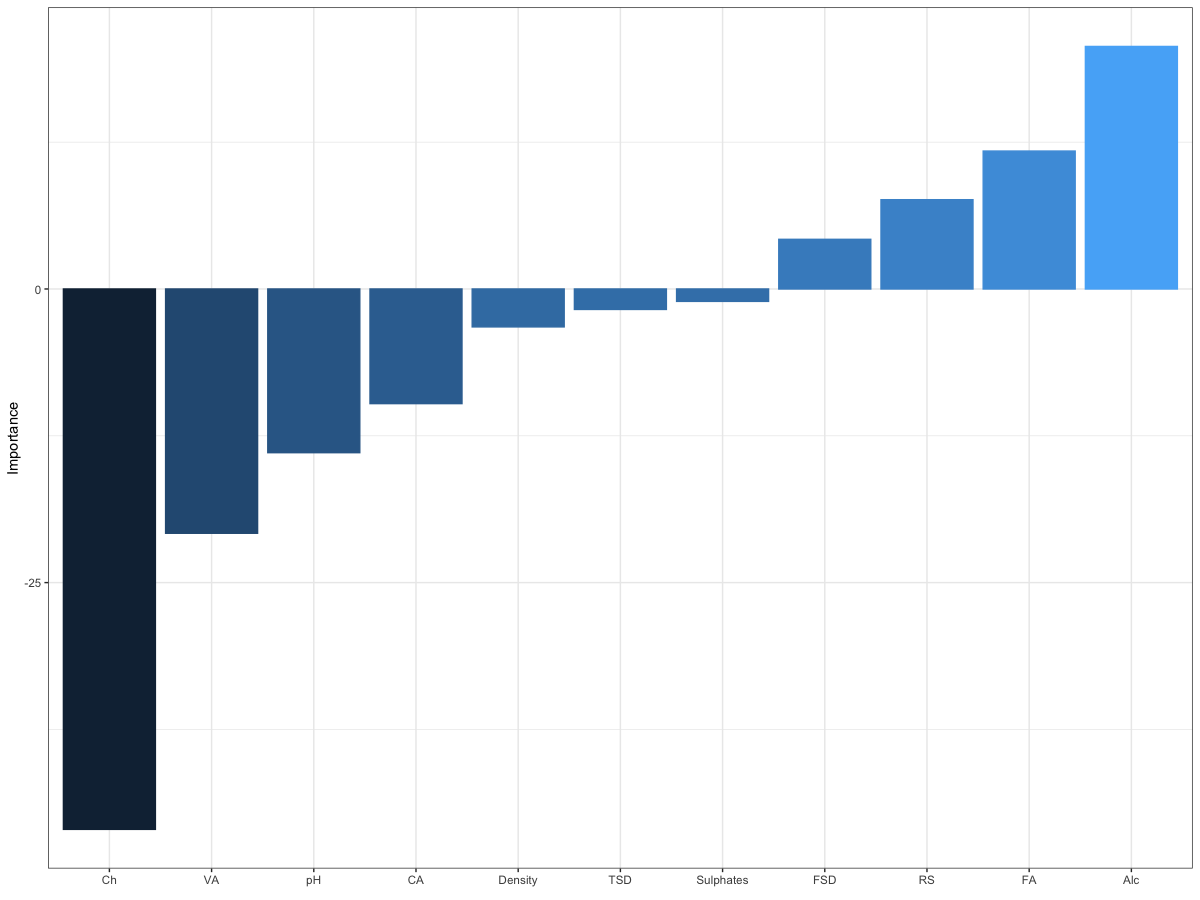


# Appendix 4 – Boosting User Loss versus Cutoff

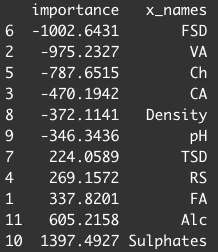


# Appendix 5 – Olden Variable Importance – Regression vs Classification

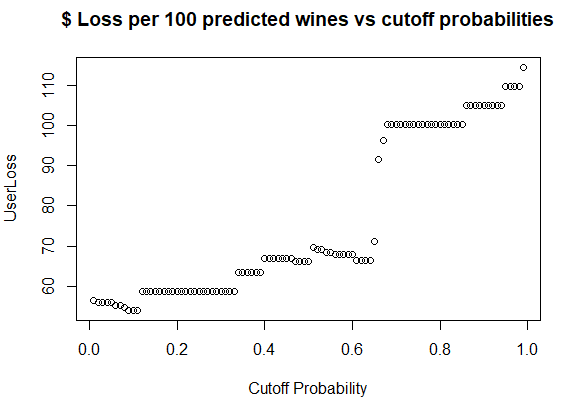
Regression Variable Importance



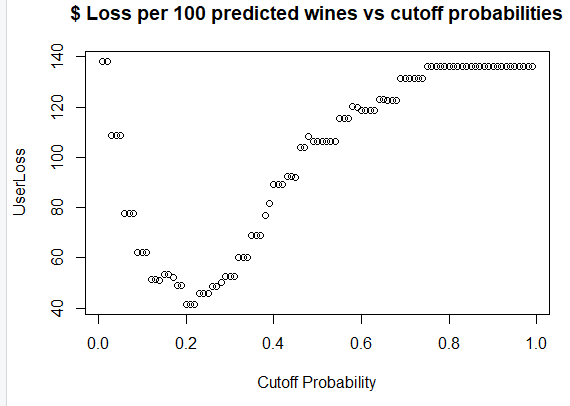
Classification Variable Importance



# Appendix 6 – NN User Loss versus Cutoff



# Appendix 7 – kNN User Loss versus Cutoff



# Appendix 8 – Random Forest Loss User Loss versus Cutoff

