Machine Learning with Applications 7 (2022) 100241

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Overcoming the ordinal imbalanced data problem by combining data processing and stacked generalizations

Marine Desprez [a,b,*], Kyle Zawada [a,b], Daniel Ramp [a]

[a] Centre for Compassionate Conservation, Faculty of Science, University of Technology Sydney, Ultimo, 2007, NSW, Australia
[b] The Yield Technology Solutions, 50 Holt St, Surry Hills, 2010, NSW, Australia

ARTICLE INFO

Keywords:
Stacked generalizations
Machine learning
Ordinal data
Imbalanced data
Random forests
Resampling methods
Rare events
Classification

ABSTRACT

Ordinal imbalanced datasets are pervasive in real world applications but remain challenging to analyse as they require specific methods to account for the ordering information and imbalanced classes. Failure to account for both those characteristics can substantially impact the model predictive performance. However, existing methods tend to focus either on ordinality or imbalance, rather than addressing both simultaneously. The few approaches that do account for both characteristics are not always easy to implement for non-advanced analysts and simpler approaches are needed to facilitate appropriate data processing. Here, we developed a general approach using some of the most popular machine learning algorithms to ensure appropriate processing of ordinal imbalanced datasets and to optimize the predictions of all classes. After transforming the multi-class ordinal problem into a well-known binary problem, we implemented several different resampling methods in a decision-tree classifier. We then used a stacked generalization algorithm to combine the classifiers to improve model predictive performance. To test our approach, we used two ordinal imbalanced datasets on student performance and wine quality. Individual resampling techniques tended to improve the accuracy of minority classes, while simultaneously increasing the number of false positives in those classes. This resulted in a decrease, sometimes substantial, in accuracy of other classes. The stacking model offered a good compromise between improvement in accuracy of minority classes and mitigation of reduced accuracy in other classes. Our approach provided useful insights into modelling strategies that should be favoured for implementation in production that involve these common datasets, depending on the end-user interests.

1. Introduction

Machine learning applications typically involve using historical and current data to detect trends and behaviours and to forecast events into the future. Generally collected through surveys and questionnaires, these data often present a natural order between classes. This characteristic makes the analysis of ordinal data challenging because class values are treated as a set of unordered categories by common machine learning classification algorithms. This mistreatment can lead to a loss of valuable information about the categories order, and potentially to poor prediction accuracy. To avoid this issue, Frank and Hall (2001) developed a simple approach to reduce an ordinal classification task to the well-known binary problem, preserving inherent ordering information without modifying the underlying learning scheme. However, integrating order information is not the only challenge in analysing ordinal data as they also typically present a strong imbalance between classes, where some classes are significantly more frequent than others. The development of approaches dealing with class imbalances and class ordinality are still rarely explored simultaneously despite the prevalence of imbalanced ordinal data across a wide range of fields, including ecology, agriculture, medicine, and social science (Kim, Kim, & Namkoong, 2016).

Conventional classification methods tend to be biased towards majority classes of imbalanced data, reducing predictive performance of minority classes (He & Garcia, 2009). However, rare events are often highly relevant to end users (e.g., fraud detection, disease diagnosis, anticipation of catastrophic events, etc.), and reliable model performance in predicting minority classes is therefore particularly critical. Because of this, classification of imbalanced datasets has attracted increasing attention in the last few years and several methods have been developed to tackle this issue (e.g., sampling strategies, cost-sensitive algorithms, hybrid/ensemble methods) (Ali, Shamsuddin, & Ralescu, 2013; Chawla, 2009; Chawla, Cieslak, Hall, & Joshi, 2008; Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2011). Resampling techniques are the

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author at: Centre for Compassionate Conservation, Faculty of Science, University of Technology Sydney, Ultimo, 2007, NSW, Australia.
E-mail addresses: marine.desprez@gmail.com (M. Desprez), kyle.zawada@uts.edu.au (K. Zawada), Daniel.ramp@uts.edu.au (D. Ramp).

https://doi.org/10.1016/j.mlwa.2021.100241
Received 25 July 2021; Received in revised form 14 December 2021; Accepted 15 December 2021
Available online 21 December 2021

most widely used approach because of their simplicity to understand and implement. These techniques work at the data level by modifying the number of instances in majority and minority classes to balance the data distribution independently of the learning algorithm (Leevy, Khoshgoftaar, Bauder, & Seliya, 2018). No single resampling technique works best for all classification problems (Kuhn, Johnson, et al., 2013; Loyola-González, Martínez-Trinidad, Carrasco-Ochoa, & García-Borroto, 2016), as their effectiveness depends on the level of imbalance in the data and the nature of the classifier used (García, Sánchez, & Mollineda, 2010). Furthermore, implementing resampling techniques may not necessarily improve prediction accuracy, as different types of resampling methods may improve predictive performance for a subset of classes under specific conditions, but may be unable to cover the entire space of the problem.

Stacked generalization, or stacking, has been increasingly used to combine several learning algorithms (called base-models) to solve the same problem (Wolpert, 1992). The idea behind this ensemble machine learning algorithm is that combining multiple models together can produce a more powerful model that achieves better predictive performance than the base-models alone. Stacking has been successfully used in both supervised and unsupervised problems and is often found at the top rankings of many machine learning competitions. Typically, base models consist of a heterogeneous collection of model types (e.g., decision trees, linear regression, support vector machine) so that the predictions produced by those models present a certain level of diversity (Witten & Frank, 2005). However, stacking different configurations of the same model type (e.g., various resampling techniques) has rarely been explored.

Imbalanced classification problems also complicate the evaluation of predictive performance. Popular classification metrics, like accuracy or classification error, assume a balanced class distribution and may be misleading when data are imbalanced (Huang & Ling, 2005; Loyola-González et al., 2016). Decision-tree classifiers for example, are built through recursive partitioning procedure that minimize the overall error regardless of the class distribution. Because of this, decision trees may report high overall accuracy while misclassifying most objects from the minority class. In that case, model accuracy mainly reflects the predictive performance in the majority class. A simple solution to accurately reflect model performance in minority and majority classes is to calculate the metric per class. Precision and recall derived from the confusion matrix provide additional use as evaluation metrics as they focus on a single class and provide information about the type of errors made in case of misclassification.

Although ordinal datasets, class imbalance, and stacking have been well studied in the literature individually, the combined problem has received little attention (but see Kerwin & Bastian, 2021 for an application on the binary classification). Ordinal imbalanced datasets are pervasive in data science and the pitfalls of mishandling classification prediction accuracy are numerous. In this study, we investigated a simple approach to improve the prediction of classes in ordinal imbalanced datasets. We used two real-world datasets from two major fields of machine learning applications: education and agriculture. The first dataset contained information about student performance and was used to predict the grade obtained by the students in the final evaluation of the year. The second dataset contained data about the physicochemical properties of wine samples and was used to predict wine quality. Classes in both datasets were ordered and imbalanced. First, we transformed the datasets to train decision-tree models while retaining the ordering information between classes. Then, we tested and compared the model predictive performance of several resampling techniques that balanced the class sample sizes in the dataset. Finally, we compared the performance of different stacking combinations of resampling methods implemented in the same learning algorithm.

2. Material and methods

2.1. Data sources

We used two ordinal imbalanced datasets that have been widely described in the literature: the student performance dataset and the wine quality dataset. Both datasets are publicly available on the UCI machine learning repository (Dua & Graff, 2017).

The student performance dataset contains data collected in Portugal on 659 students during the 2005–2006 school year. The objective of the analysis was to predict the grade obtained by the students in the third and final evaluation of the year. The 32 predicting variables included in the dataset described the demographic, social, and school related environment of each student. A full description of the data is available in Cortez and Silva (2008). All nominal features were encoded into a set of binary variables. Following a common honours grading system associated with the 20-point grading scale of the response variable, students were grouped in six levels: (1) 18 to 20: Highly honourable; (2) 16 to 17.99: Highest honours; (3) 14 to 15.99: High honours (4) 12 to 13.99: Satisfactory (5) 10 to 11.99: Sufficient and (6) 0 to 9.99: Fail. Most students were classified as levels 4 and 5 (Fig. 1).

The wine quality dataset contains 11 variables related to the physicochemical properties of 4,898 samples of the white variant of Vinho Verde, a wine produced in the northwest region of Portugal. Each sample was also evaluated by sensory assessors that graded the wine quality in a scale ranging from 0 (very bad) to 10 (very good). The number of samples assigned to each quality grade was highly imbalanced and no samples were assigned the grades 0, 1, 2 and 10 (Fig. 2). A full description of the data is available in Cortez, Cerdeira, Almeida, Matos, and Reis (2009).

2.2. Binary decomposition

Analyses of the student performance and wine quality datasets followed the same general approach but were conducted separately. We applied the method described by Frank and Hall (2001) to make use of the ordering information contained in the response variables, while using a decision tree learner as the classifier (Fig. 3). First, student grades and wine quality grades were transformed into a set of binary classes (e.g., grade >1: yes or no; grade >2: yes or no, etc.) and a new dataset was derived by combining the new binary class with the predictor variables. A decision tree learner was then run on each dataset separately and probabilities for each binary class were estimated. Finally, probabilities for original grades were calculated from the probability estimates of each binary class (e.g., $Pr_{(grade = 2)} = Pr_{(grade>1)} - Pr_{(grade>2)}$). The predicted class was the one with the maximum associated probability. In this study, Random Forests was used as the decision tree classifier. However, the use of different classifier (XGBoost, Neural net, etc.) is possible.

A test set containing 20% of observations from the original dataset was put aside before model training. The training set was randomly split into 3 folds to conduct cross-validation and hyper-parameter tuning. Because the focus of this study was not to identify the best predictors of student performance or wine quality, we did not conduct feature selection on the predictors contained in the datasets.

The analysis was conducted in R 4.0.1 (R Core Team, 2021) and the 'Ranger' package (Wright & Ziegler, 0000) was used for training the random forests models.

2.3. Sampling methods

The class distribution in both datasets was imbalanced (Figs. 1 and 2). We implemented and tested several data sampling methods to balance the distribution of the classes and improve the predictive performance of the model. Many different sampling techniques exist but we choose to focus on some of the most popular and easiest to

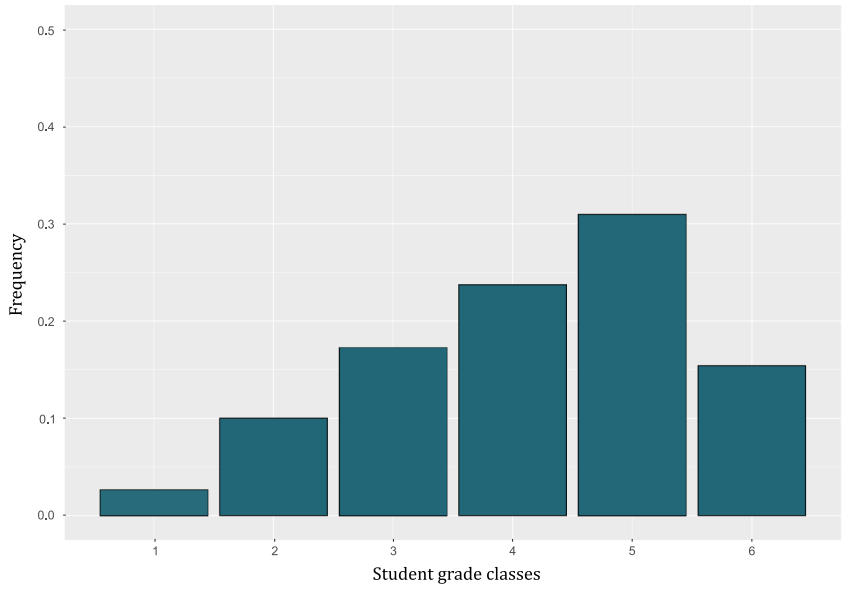

Fig. 1. Students frequency per grading class.

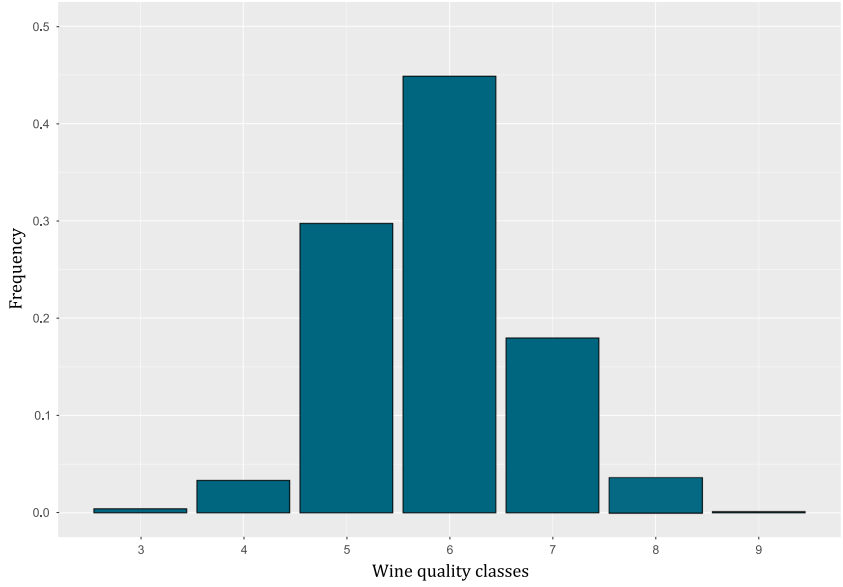

Fig. 2. Wine samples frequency per quality class.

implement. However, the approach described in this study can be modified to implement different sampling methods. The sampling techniques tested were:

- The over-sampling method: randomly duplicate examples from the minority class. By replicating copies of the minority class examples, this method may increase the risk of overfitting.
- The under-sampling method: randomly delete samples from the majority class. One major limitation of this method is that a vast amount of data may be discarded and the subsequent information loss can result in a decrease in model performance.
- The combined method: randomly resample the training set by deleting instances from the majority class and duplicating observations from the minority class.
- The SMOTE method (Chawla, Bowyer, Hall, & Kegelmeyer, 2002): new instances of the minority class are artificially generated using the nearest neighbours of existing points. A line is drawn between neighbouring instances and a point along that line is picked as a new record for the minority class. Observations from the majority class are simultaneously randomly under-sampled leading to more even sample sizes between classes.

The selected sampling method was applied during the cross-validation process on the training set only. Results were compared to the ones obtained from a model without a sampling method applied. We used the ovun.sample function from the R package 'ROSE' (Lunardon, Menardi, & Torelli, 2014) to conduct the over, under and combined sampling of the dataset while the 'smotefamily' package (Siriseriwan, 2019) was used to perform the SMOTE sampling.

2.4. Stacking

Stacking (or stacked generalization) is an ensemble algorithm that combines the predictions from multiple machine learning models fitted on the same dataset (Wolpert, 1992). The architecture of a stacking model involves two or more base models and a meta-model that

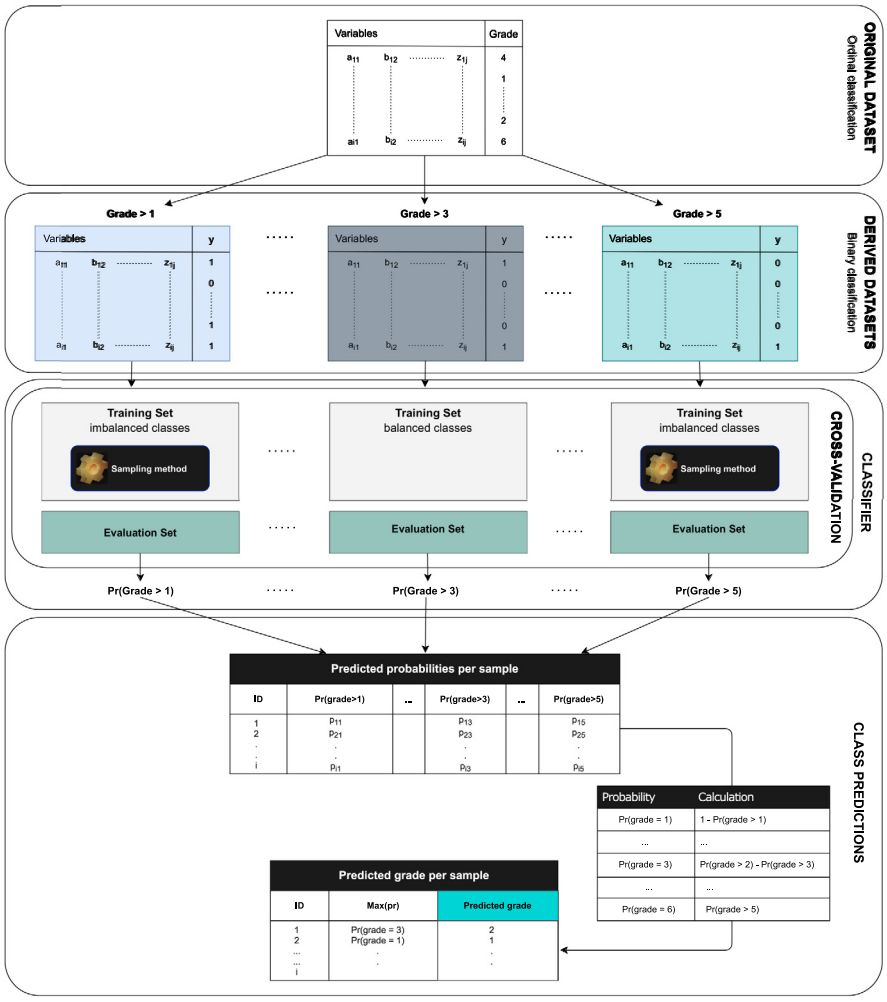

Fig. 3. Modelling framework from data pre-processing to classification.

combines the predictions of all the base models (Figs. 4 and 5). The base models used in our study were random forest models including different sampling techniques (see Sections 2.2 and 2.3 for details). The predictions from those random forest models were then used as the input for the meta-model. Two different types of predictions were tested as input: (i) the predicted grades (Fig. 4) and (ii) the probability values (Fig. 5). When quality grades were used as the input, the meta-learner was an ordinal logistic regression. In the case of the probability values, we used a logistic regression as the meta-model. These meta-learners were chosen so that the ordering information included in the response variables was preserved. All combination of base models were tested (see appendix A) and only the model with the best predictive performance (see Section 2.5 below) were presented in the results.

Ordinal logistic regressions were built using the R package 'ordinal' (Christensen, 2019), while logistic regressions were built using the glm function.

2.5. Measures of model predictive performance

In the case of ordinal imbalance data, evaluation metrics calculated across all classes may not be fully robust as they mainly reflect the model performance in the majority classes. To accurately reflect the predictive performance of the model in all classes, each metric was also reported per class.

We evaluated the model predictive performance by calculating the percentage of correct classification (PCC), which is the number of instances correctly identified divided by the total number of observations in the class of interest. Considered on its own, this measure may be misleading, so precision and recall were also calculated. Precision measures the percentage of relevant predictions made by the model, defined as the number of true positives (i.e., a sample that was correctly classified by the model) divided by the sum of true positives and false positives (i.e., a sample that was wrongly classified into the class of interest). This metric is particularly important in studies where false positives should be minimized (e.g., in spam detection studies, important emails should not be wrongly classified as spam). Recall, on the other hand, measures the ability of a model to find all the data points of interest within a dataset. It is calculated by dividing the number of true positives by the sum of the true positives and false negatives (i.e., a sample not assigned by the model to its correct class). This measure is critical in studies like disease diagnosis where it is paramount to correctly identify all sick patients. A perfect classifier has both precision and recall equal to 100%. We presented in the results the models with the highest average PCC, precision, and recall.

Another commonly used estimator of the predictive performance of an ordinal classifier is the magnitude of the error between predictions and actuals. MAE (mean absolute error) and the RMSE (root mean squared error) increase with the difference between the actual and predicted classes, meaning that misclassifications are not estimated as equally costly. Since the errors are squared in the calculation of the RMSE, the RMSE penalizes large prediction errors more than the MAE. However, MAE is easier to interpret than RMSE. We calculated both

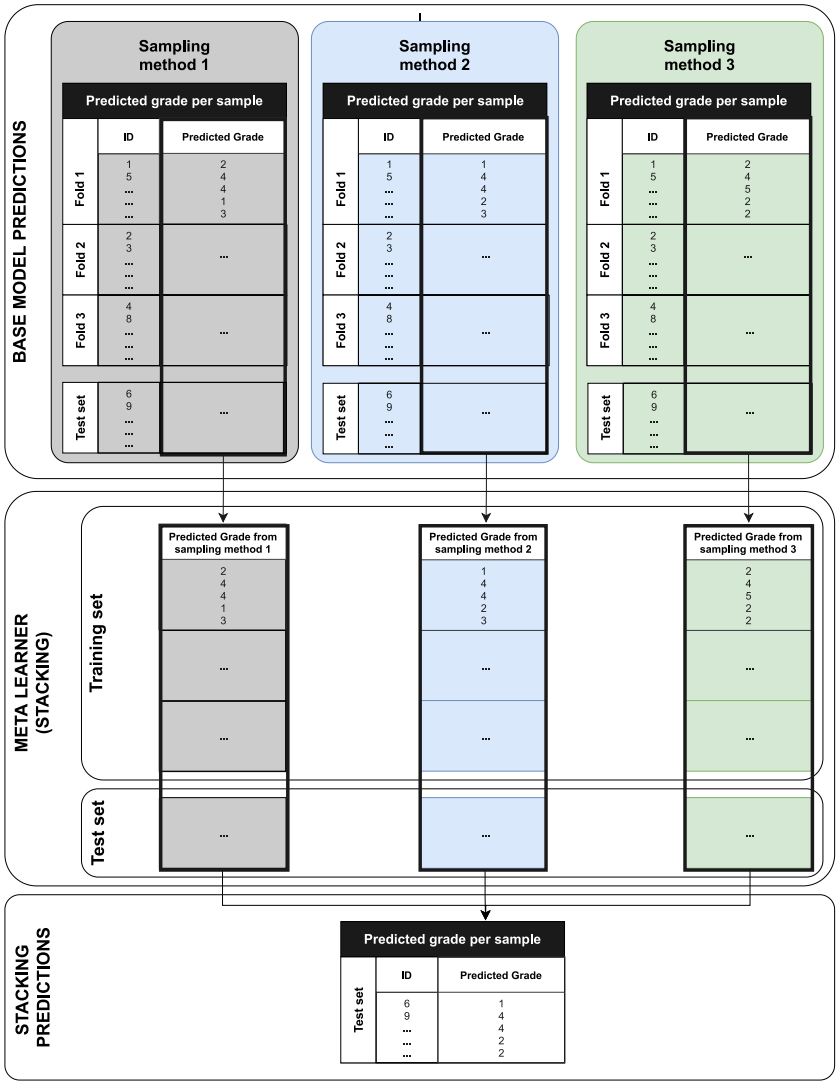

Fig. 4. Overview of the stacking framework with ordinal logistic regression as meta-learner.

MAE and RMSE. We did not discuss these values in the manuscript but reported them in the Appendix B.

3. Results

3.1. Student performance

3.1.1. Comparison of sampling models

The SMOTE model was the only sampling model that improved the average PCC compared to the no sampling model (75.38% vs. 73.85% respectively). No model achieved optimal predictive performance for all grade classes (Fig. 6). The PCC of the minority class 1 was constant across all models (66.67%), except in the under-sampling model where it reached 100%.

The combined sampling model had the highest PCC in the classes 2, 4, and 6 (Fig. 6). This was explained by a lower number of false negatives in those classes (i.e., less students were wrongly assigned to other classes; higher recall). However, this gain in PCC was also associated with an increase in the number of false positives (i.e., more students from other classes were wrongly classified in classes 2, 4, and 6; lower precision). This decrease in precision of the classes 2, 4, and 6 led to a significant drop in PCC and recall (i.e., higher number of false negatives) in the classes 3 and 5, confirming that a significant number of students from those classes were wrongly assigned to the classes 2, 4, and 6 (Fig. 6). The same tendency was observed in the over-sampling and under-sampling models. Higher PCC in the minority grade classes was consistently associated with a higher recall (i.e., less false negatives) but lower precision (i.e., more false positives). This led to a decrease in PCC and recall (i.e., increase of false negatives) in the other classes, with students being wrongly classified into the minority classes. This was particularly noticeable in the under-sampling model. This model was the only one showing an increase in PCC and recall in minority class 1 (Fig. 6). The PCC of class 6 was also substantially higher in this model. However, this gain in PCC came at the cost of the PCC and recall of classes 3 and 5, which dropped to an extreme 0% for class 3 (compared to 52.94% in the no sampling model) and 39.02% for class 5 (compared to 95.12% in the no sampling model).

The SMOTE model was the only sampling model that did not show a decrease in average precision and recall compared to the no sampling model. PCC of classes 4 and 6 were higher in the SMOTE model than in the no sampling one. This gain in PCC was again due to a lower number of false negatives; but contrary to the other sampling models, precision remained high (i.e., number of false positives did not increase). The PCC of classes 2 and 5 were lower than in the no sampling model while it was similar for the rest of the classes.

3.1.2. Stacking model

The stacking model with the best average PCC, precision, and recall combined the no sampling model, the combined sampling model, the

Fig. 5. Overview of the stacking approach with logistic regression as meta-learner.

under-sampling model, and the SMOTE model as base models. The grade classes predicted from each of those models were used as the input into the ordinal logistic regression (see Appendix A).

The stacking model had the highest average precision and recall of all models, and its average PCC was equal to the PCC of the SMOTE model. The PCC per class was higher or equal to the one obtained in the SMOTE model, except for class 5 (90.24% in the stacking model compared to 92.64% in the smote model). The PCC of class 1 remained the same (66.67%) in the stacking model as in most other sampling models.

The stacking model had a lower PCC in minority classes 2 and 6 than the combined sampling, over-sampling, or under-sampling models (Fig. 6). However, the precision of those classes was higher in the stacking model, indicating a limited number a false positives. The drop in PCC in classes 3 and 5 was also more limited in the stacking model.

3.2. Wine quality

3.2.1. Comparison of sampling models

The model with the highest average PCC (68.47%) was the model that did not implement any sampling method (Fig. 7), and no model achieved optimal predictive performance for all grade classes. The implementation of sampling methods in the model tended to improve the PCC of classes 4, 7, and 8, except in the case of the under-sampling model (Fig. 7). Like the analysis of student performance, the gain in PCC in those classes came at the cost of the PCC and recall in other classes. More specifically, a higher PCC in the classes 4, 7, and 8 was associated with higher recall but lower precision (i.e., higher number of samples wrongly classified) into the classes 4, 7, and 8. This led to the increase of false negatives in classes 5 and/or 6, and in turn a drop in their PCC (Fig. 7). The highest PCC for classes 4, 7, and 8 was achieved in the combined sampling model.

The under-sampling model was the only model that predicted minority classes 3 and 9 with a high PCC (Fig. 7). However, the increase of false positives in those classes was substantial and the drop in PCC of other classes too high for this model to be considered useful. The PCC of class 9 was null in all the other models. PCC of class 3 was also null or very low (12.5% in the SMOTE model) in all other models (Fig. 7).

3.2.2. Stacking model

The stacking model with the highest average PCC, precision, and recall combined the no sampling model and the combined sampling model as base models (see Appendix A). The probability values from each model were used as the input into a logistic regression model to predict the wine quality grade of each wine sample.

The stacking model had the best average PCC and precision of all the models tested (Fig. 7). Compared to the no sampling model (second best average model), the stacking model had a better PCC in all the classes except the majority class 6. The loss of PCC in class 6 was limited in the stacking model compared to the other sampling models (Fig. 7). The PCC of class 9, the highest quality grade, remained null in the stacking model.

Except for the under-sampling model, all the sampling models had a higher or similar PCC in the good quality wine classes (7 and 8) than the stacking model (Fig. 7). However, the precision associated to those classes was much lower in the sampling models than in the stacking one. This indicated that the sampling models wrongly classified a high number a wine samples into the good quality classes, while the stacking model tended to have a higher number of false negatives. The PCC of

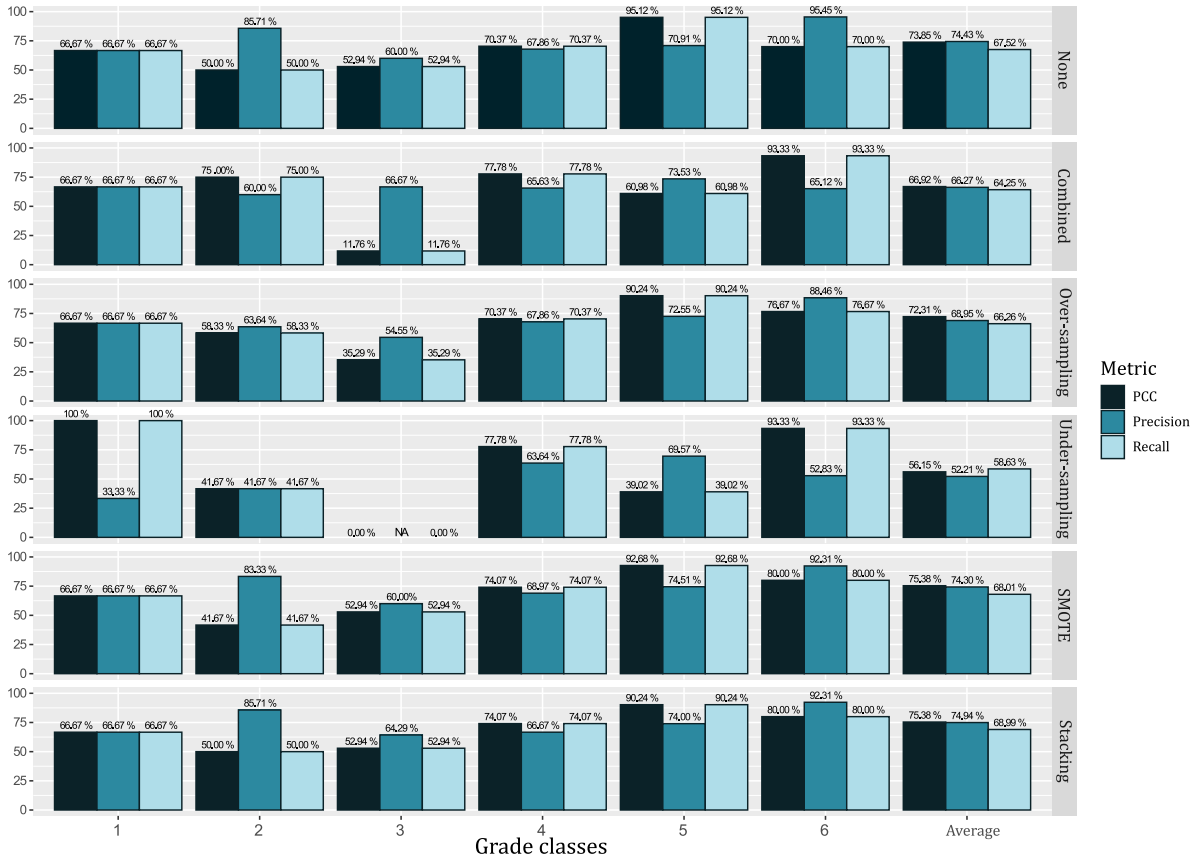

Fig. 6. Percentage of correct classification, precision and recall for the model without resampling method, the models with combined, over and under resampling method and the best stacking model for the student performance dataset.

the worse quality wine class remained very low (12.5%) in the stacking model. The PCC was equal to the PCC obtained in the SMOTE model but the precision in the stacking model was much higher (100% vs. 33.33%).

4. Discussion

Ordinal imbalanced data are prevalent in real-world applications, but their analysis requires the use of specific methods to account for the ordering information and to balance the class distribution (Kim et al., 2016). In this paper, we investigated a novel and simple approach to optimize the class predictions in two ordinal imbalanced datasets. We used a stacked generalization for classification after integrating the ordering information and the most popular sampling methods into a decision-tree model.

The individual sampling techniques we tested were not optimal at predicting all grade classes. Prediction accuracy for minority classes were generally improved when resampling methods were implemented in the random forests. However, it was also consistently associated with a decrease in the percentage of correct classification (PCC) of other classes. This was due to a lower precision (i.e., increase of false positives) in the minority classes, which led to a lower recall (i.e., increase of false negatives) in other classes. Costs associated with the improvement of the minority class accuracy was also reported in binary classification problems by García et al. (2010), who noticed that resampling methods increased the true positive rate in minority class but simultaneously decreased the true negative rate. Those costs were limited in the stacking model that tended to optimize PCC, precision, and recall in all classes. Therefore, the gain in accuracy of minority classes was lower than in some of the resampling models, but the loss of precision in minority classes and decrease in PCC and recall in majority classes was minimized, providing a good overall compromise.

The choice of predictive model ultimately depends upon end-user interest in specific classes. In the wine quality analysis, the stacking model optimized the predictive performance of all quality classes. However, viticulturists and wine producers are most likely to be interested in predicting which samples produce the best wine quality. None of the models were able to accurately predict the best wine quality class, except for the under-sampling model which was uninformative because of the massive loss of precision in the minority class. For the second-best wine quality class (i.e., class 8), two different strategies could be adopted. First, it may be decided to choose a predictive model that limits the risks of wrongly classify samples as good quality ones (i.e., limiting the number of false positives) despite a slightly higher risk of misclassifying good samples in lower quality categories (i.e., higher number of false negatives). In this case, the stacking model is the appropriate choice as it provided the highest precision in class 8 and a PCC only slightly lower than the highest one obtained by the combined resampling model. An alternative strategy may be to limit the risk of misclassifying a good sample in lower categories (i.e., limiting the number of false negatives) as much as possible, even if it leads to classifying lower quality samples in the high categories (i.e., higher number of false positives). This strategy would favour the use of the combined sampling model.

In the student performance analysis, the SMOTE model had very similar results to the stacking model. The only important differences were between class 2, which had higher PCC, precision, and recall in the stacking model, and class 5, which was marginally better predicted by the SMOTE model. Like the wine quality analysis, the choice of models ultimately depend on end-user interests and objectives. Both stacking and SMOTE models would be an appropriate choice if the study aim was to predict the students in each category with the highest possible PCC, precision, and recall. Schools may also have a particular interest in predicting which students will fail, or at risk of failing, the

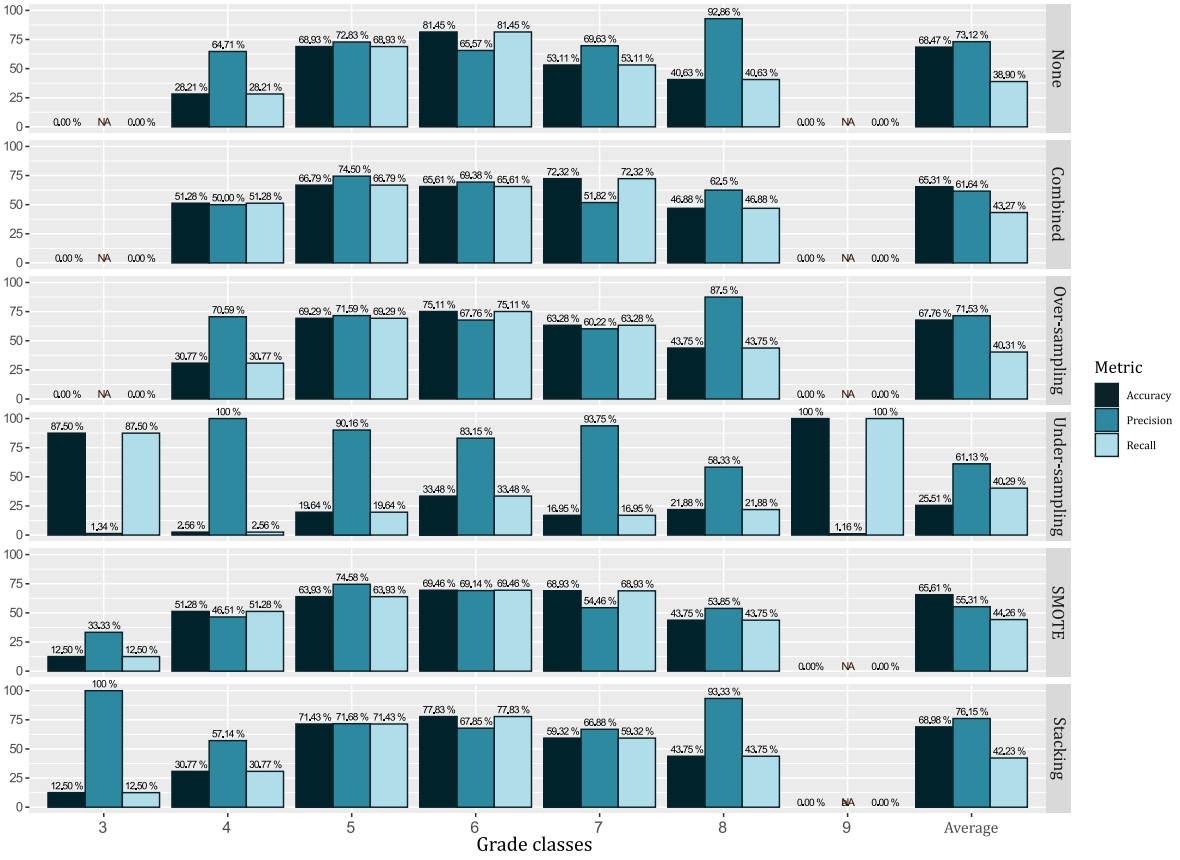

Fig. 7. Percentage of correct classification, precision and recall for the model without resampling method, the models with combined, over and under resampling method and the best stacking model for the wine quality dataset.

final evaluation of the year (classes 6 and 5 respectively), to provide those students with tutoring support system (Cortez & Silva, 2008). The primary goal may be to reduce the risk of missing students of interest, even if that means selecting students that are not actually in need of support. In that case, the combined model that limits the number of false negatives predicted in classes 6 would be favoured. In contrast, some schools may choose to favour models that limit the number of false positives predicted. This may be the strategy of a school that cannot afford to support students that are not in need, due to limited resources, even if that means missing a few students of interest. In that case, the SMOTE or stacking models may be more appropriate.

Most sampling methods were unable to make accurate predictions when the number of instances in a class was too low. For example, the PCC for the best wine quality class (i.e., class 9, which only contained 5 samples) was null in most models. The under-sampling model was the only one that classified samples into class 9 with a perfect accuracy, but this came at the cost of substantial loss of accuracy in other classes. If the focus of analysis is to predict the best wine quality samples, with no interest in other wine quality categories, then an under-sampling model would be an appropriate choice. However, particular attention should be paid to the per class precision and recall obtained by this model. The precision of the best wine quality class was almost null (1.16%, very high number of false positives), and the recall in all other classes was also extremely low. This indicates that the model assigned the top wine quality class to most samples from the dataset. Such a model would be uninformative to most end-users. A minimum number of instances per class is required to make reliable predictions. In the case of a class with a too small sample size, it may be more pragmatic to group samples from these classes into adjacent classes. Our findings highlight the importance and utility of considering additional metrics like precision and recall when evaluating the predictive performance of a model.

Ordinal imbalanced data are challenging to analyse and require methods that simultaneously account for the order information and imbalance between classes. We applied a general approach that ensured the appropriate processing of ordinal imbalanced data and optimized classes prediction in student performance and wine quality. Individual resampling methods produced different results and while they improved the accuracy of minority classes, associated costs in precision and accuracy of other classes were sometimes substantial. Using stacking to combine several resampling techniques, a good compromise between the accuracy improvement of minority classes and the minimization of associated costs in other classes was achieved. However, individual resampling techniques may still be favoured over the stacking model, depending on the end-user interests in specific classes. Generally, stacking models tended to limit the number of false positives produced by the model and may therefore not be the appropriate choice in studies favouring false positives over false negatives. Computational time and amount of data available are also important elements to consider when choosing the type of predictive model to implement in production. Stacking models can be computationally expensive to run and also require a subset of the dataset to be put aside for final testing. These data cannot be used for the training of base models, which may be an issue in ordinal datasets with categories of very small sample sizes, especially if cross-validation is conducted during the training process. Therefore, the choice of whether to implement a stacking model in production need to be weighed against the predictive gains and associated computational costs. Overall, our approach provided useful transparency and insight into modelling results and can be used to facilitate the choice of an appropriate model for overcoming the inherent challenges that ordinal imbalanced data create.

CRediT authorship contribution statement

Marine Desprez: Conceptualization, Methodology, Data curation, Formal analysis, Visualization, Writing – original draft. **Kyle Zawada:** Validation, Writing – review & editing. **Daniel Ramp:** Supervision, Funding acquisition, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project was funded by the Food Agility CRC, Australia project FA019, titled "Harvest timing and yield prediction" in partnership with The Yield. We thank Dr Evan Webster and Dr Boyu Ji for fruitful discussions around the development of the modelling framework, and Ashley Rootsey and Dr Kate May for facilitating the collaboration between The Yield and UTS.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mlwa.2021.100241.

References

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem. *International Journal of Advances in Soft Computing and its Applications*, *5*, 176–204.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In R. L. Maimon O. (Ed.), *Data mining and knowledge discovery handbook* (pp. 875–886). Boston, MA: Springer.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *17*, 225–252.

Chawla, N. V., Cieslak, D. A., Hall, L. O., & Joshi, A. (2008). Automatically countering imbalance and its empirical relationship to cost. *Data Mining and Knowledge Discovery*, *17*, 225–252.

Christensen, R. H. B. (2019). *Ordinal — Regression models for ordinal data*. R Package Version 2019.12-10. URL: https://CRAN.R-project.org/package=ordinal.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*, 547–553.

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *Proceedings of 5th annual future business technology conference, port* (pp. 5–12). EUROSIS-ETI.

Dua, D., & Graff, C. (2017). UCI machine learning repository. URL: http://archive.ics.uci.edu/ml.

Frank, E., & Hall, M. (2001). A simple approach to ordinal classification. In F. P. De Raedt L. (Ed.), *European conference on machine learning* (pp. 145–156). Berlin, Heidelberg: Springer.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*, 463–484.

García, V., Sánchez, J. S., & Mollineda, R. A. (2010). Exploring the performance of resampling strategies for the class imbalance problem. In *International conference on industrial, engineering and other applications of applied intelligent systems* (pp. 541–549). Berlin, Heidelberg: Springer.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *121*, 1263–1284.

Huang, J., & Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*, 299–310.

Kerwin, K. R., & Bastian, N. D. (2021). Stacked generalizations in imbalanced fraud data sets using resampling methods. *The Journal of Defense Modeling and Simulation*, *18*, 175–192.

Kim, S., Kim, H., & Namkoong, Y. (2016). Ordinal classification of imbalanced data with application in emergency and disaster information services. *IEEE Intelligent Systems*, *31*, 50–56.

Kuhn, M., Johnson, K., et al., New York (2013). *Applied predictive modeling*. Springer.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, *5*, 1–30.

Loyola-González, O., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., & García-Borroto, M. (2016). Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing*, *175*, 935–947.

Lunardon, N., Menardi, G., & Torelli, N. (2014). Rose: A package for binary imbalanced learning. *R Journal*, *6*, 82–92.

R Core Team (2021). *R: A language and environment for statistical computing*. Austria: R Foundation for Statistical Computing Vienna, URL: https://www.R-project.org/.

Siriseriwan, W. (2019). Smotefamily: A collection of oversampling techniques for class imbalance problem based on SMOTE. URL: https://CRAN.R-project.org/package=smotefamily R package version 1.3.1.

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques* (second ed.). San Francisco: Morgan Kaufmann Publishers.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*, 241–259.

Wright, M. N., & Ziegler, A. (0000). ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software, 77.