

Predicting How Much a Consumer Is Willing to Pay for a Bottle of Wine: Dealing With Data Imbalance

Hugo Alonso^{1,2}^a and Teresa Candeias¹^b

¹Universidade Lusófona – Centro Universitário do Porto, Rua Augusto Rosa, n.º 24, 4000-098 Porto, Portugal

²Universidade de Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

Keywords: Wine, Classification, Data Imbalance, Re-Sampling, Learning Methods, Predictive Models.


Abstract: The wine industry has becoming increasingly important worldwide and is one of the most significant industries in Portugal. In a previous paper, the problem of predicting how much a Portuguese consumer is willing to pay for a bottle of wine was considered for the first time ever. The problem was treated as a multi-class ordinal classification task. Although we achieved good prediction results, globally speaking, it was difficult to identify rare cases of consumers who are interested in paying for more expensive wines. We found that this was a direct consequence of data imbalance. Therefore, here, we present a first attempt to deal with this issue, based on the use of re-sampling strategies to balance the training data, namely random under-sampling, random over-sampling with replacement and the synthetic minority over-sampling technique. We consider several learning methods and develop various predictive models. A comparative study is carried out and its results highlight the importance of a careful choice of the re-sampling strategy and the learning method in order to get the best possible prediction results.


1 INTRODUCTION

Wine market became more demanding with the growing number of new global players and a changing consumer behavior. With the heterogeneity of wine markets, several studies suggested the use of segmentation methodologies to understand wine consumer behavior (Bruwer et al., 2002; Thach and Olsen, 2006; Kolyesnikova et al., 2008; Koksál, 2021; Payini et al., 2022). With thousands of wine brands, styles and regions, consumers are frequently confused when purchasing wine. According to (Rouzet and Seguin, 2004), in order to match wine consumers' preferences with wine characteristics, segmentation divides markets that can be reach with different marketing instruments. Usually, the marketing segmentation variables are geographic, demographic, psychographic and behavioral (Kotler and Keller, 2006). Segmentation based on lifestyle has also been applied in the US, although with the purpose to underline motivations and occasions of consumption (Thach and Olsen, 2005). Overall, an effective marketing strategy is required and, in this context, understanding wine consumers'

needs and buying habits plays an important role in market segmentation.

In a previous paper (Alonso and Candeias, 2022), the problem of predicting how much a Portuguese consumer is willing to pay for a bottle of wine was considered for the first time ever. More precisely, given information about an individual, such as his/her age and income, we were interested in predicting how much he/she is willing to spend in a bottle: less than EUR 2.99; between EUR 3 and 4.99; between EUR 5 and 9.99; EUR 10 or more. Since these intervals can be viewed as ordered classes, the prediction problem was treated as a multi-class ordinal classification task. Using several types of predictive models and learning methods, we achieved good results in terms of the overall accuracy and r_{int} (a measure of association between the ordinal variables true class and predicted class) (Pinto da Costa et al., 2008; Pinto da Costa et al., 2014). However, we found that all classifiers had more difficulty in correctly predicting cases from higher classes and that this was related to our data imbalance. Note that, since most people are willing to pay less and only a small number of people are willing to pay more for a bottle of wine, lower classes are much more frequent than higher ones. In this context, identifying consumers who are willing to

^a <https://orcid.org/0000-0002-1599-5392>

^b <https://orcid.org/0000-0002-3371-9869>

pay more for a bottle of wine corresponds to predicting rare events. Here, we present our first attempt to deal with this issue. Our goal is to obtain more balanced classifiers, *i.e.*, with an improved ability to predict infrequent cases without seriously compromising the prediction of frequent ones.

It is well known that the ability to predict rare events remains one of the most challenging tasks to solve in machine learning (Arafat et al., 2019). According to this reference and also to (Sun et al., 2009; Haixiang et al., 2017; More and Rana, 2021), a common strategy to cope with this problem consists in applying re-sampling methods to balance the training data. Another possibility consists, for instance, in assigning different classification costs to different classes, but deciding those costs is a difficult task and incorporating them in some data mining algorithms is not easy. In this paper, we apply three popular re-sampling techniques: random under-sampling (RUS), random over-sampling with replacement (ROSWR) and the synthetic minority over-sampling technique (SMOTE) (Ganganwar, 2012; Kotsiantis et al., 2006; Chawla et al., 2002). RUS randomly removes examples from the most represented classes in the training set, but, by doing so, it can discard potentially useful data that could be important for the induction process, thus leading to underfitting. In turn, ROSWR and SMOTE randomly add examples to the least represented classes in the training set, though in different ways: while ROSWR repeats existing examples, SMOTE generates new artificial ones. The main drawback of ROSWR is that it can increase the likelihood of occurring overfitting, because repeating examples makes them more important during the training phase. This problem is avoided by SMOTE.

Selecting proper evaluation metrics plays a key role in the task of correctly handling data imbalance. In (Branco et al., 2016), the authors survey several metrics and discuss their advantages and disadvantages. For a multi-class classification problem, like ours, they conclude that the so-called MF_1 score and related measures are suitable for performance assessment. Hence, we use them in this study.

In this work, we wish to compare our previous results in (Alonso and Candeias, 2022) with news ones we obtained by applying re-sampling strategies to balance our data set. Therefore, the remainder of the paper is organized as follows. The next section describes the data we considered. The re-sampling strategies are presented in Section 3 and the predictive models and learning methods in Section 4. The issue of how to assess performance in an imbalanced problem is addressed in Section 5 and we define suitable metrics for that purpose. Finally, the results are

shown and compared in Section 6 and the conclusions and future work are given in Section 7.

2 DATA

The data set considered in this study is the one we introduced in our previous paper (Alonso and Candeias, 2022). It has a total of 228 instances and 9 attributes. There are 8 predictive attributes or input variables, nominal and ordinal, corresponding to consumers' characteristics: gender, age, marital status, education level, region of residence, income, wine knowledge and consumption frequency. The target attribute or output variable is ordinal and corresponds to the bottle price class. The full data set is partitioned into training and test subsets, with 2/3 and 1/3 of all available instances, respectively. The partitioning is stratified and so the *a priori* class distribution is roughly the same in the three sets. The classes are $C_1 =]0, 2.99[$ euros, $C_2 = [3, 4.99[$ euros, $C_3 = [5, 9.99[$ euros and $C_4 = [10, +\infty[$ euros and their relative and absolute frequencies in the three sets are given in Table 1. Remark that the data are imbalanced: the distribution is skewed to the right, with the two lower classes being much more frequent than the two higher ones. The reason is that most people are willing to pay less and only a small number of people are willing to pay more for a bottle of wine. Further details about the data can be found in (Alonso and Candeias, 2022).

3 RE-SAMPLING STRATEGIES

Re-sampling strategies are used to balance an imbalanced training set like ours. In the following, we briefly describe three popular techniques: random under-sampling, random over-sampling with replacement and the synthetic minority over-sampling technique (Ganganwar, 2012; Kotsiantis et al., 2006; Chawla et al., 2002).

Random under-sampling consists in withdrawing from the training set instances randomly chosen from the most frequent classes, potentially until all classes have the same number of cases or roughly the same.

In turn, random over-sampling with replacement consists in adding to the training set instances randomly chosen from the least frequent classes, potentially until all classes have the same number of cases or roughly the same. Note that, when an instance is added to the augmented training set, it is always drawn with replacement from the initial training set.

Finally, just like random over-sampling with replacement, the synthetic minority over-sampling tech-

Table 1: Frequency distribution of the bottle price class variable in the full, training and test data sets.

| | | Bottle price class | | | | Total |
|-------------------------|--------------|--------------------|-------|-------|-------|-------|
| | | C_1 | C_2 | C_3 | C_4 | |
| Percentage of instances | | 38% | 42% | 16% | 4% | 100% |
| Number of instances | Full set | 86 | 95 | 37 | 10 | 228 |
| | Training set | 57 | 63 | 25 | 7 | 152 |
| | Test set | 29 | 32 | 12 | 3 | 76 |

nique adds instances to the least frequent classes. However, the way how it does it is different, as will be explained later on. Meanwhile, it should be said that the authors of the method proposed three versions of it: SMOTE, for the case where all predictive attributes are (quantitative) continuous; SMOTE-NC, for the case where there is a mixture of nominal and continuous predictive attributes; SMOTE-N, for the case where all predictive attributes are nominal. Here, since our data set has a mixture of nominal and ordinal predictive attributes, we treat ordinal variables as if they were nominal and consider the SMOTE-N version, which we describe next, as well as the way how we applied it.

In SMOTE-N, in order to add a new instance $\tilde{\mathbf{x}}$ to class C , we start by selecting an instance \mathbf{x} in C from the initial training set and determine its k nearest neighbors $\mathbf{x}_1, \dots, \mathbf{x}_k$ in C from such set. The nearest neighbors are computed using the modified version of the Value Difference Metric (Stanfill and Waltz, 1986) proposed by (Cost and Salzberg, 1993) and incorporating the suggestions in (Chawla et al., 2002). Then, a new instance $\tilde{\mathbf{x}}$ we add to the augmented training set is given by a vector whose i -th component is the most frequent value among the values of the i -th components of \mathbf{x} and k' of the k neighbors $\mathbf{x}_1, \dots, \mathbf{x}_k$, where $1 \leq k' \leq k$. In this paper, we take $k = 5$. Remark that the value of k must be lower than the number of instances in the least represented class in our case, *i.e.*, 7 (see Table 1). Then, for each possible choice of \mathbf{x} in a class C for which we want to add instances, we generate as many new instances as possible and necessary in the following way: the neighbors $\mathbf{x}_1, \dots, \mathbf{x}_k$ are split into two sets with $\lfloor k/2 \rfloor$ and $k - \lfloor k/2 \rfloor$ elements; new instances $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ are generated by combining \mathbf{x} with the neighbors in the first and second sets, respectively; the process of splitting and generation is repeated for all possible splits.

4 PREDICTIVE MODELS AND LEARNING METHODS

As was mentioned earlier, we wish to compare our previous results in (Alonso and Candeias, 2022) with

the news ones we obtained by applying re-sampling strategies to balance our data set. For this reason, we consider the same predictive models and learning methods.

Hence, we consider here three types of predictive models: artificial neural networks, support vector machines and decisions trees (Hastie et al., 2009). Two advantages of decision trees are their interpretability and the ease with which they deal with qualitative predictive variables. Artificial neural networks and support vector machines are not as easily interpretable, but very often they have better generalization results. Details about these models are given in the previous reference.

Regarding the learning methods, we consider the conventional approach to supervised classification, where the order relation between the classes is not taken into account (Hastie et al., 2009), and two ordinal supervised classification approaches, namely the so-called unimodal binomial model (Pinto da Costa et al., 2008) and a modification of Frank and Hall's method (Frank and Hall, 2001), proposed in (Cardoso and Pinto da Costa, 2007). These two ordinal learning methods and the way how they are applied to our problem are briefly described next.

4.1 The Unimodal Binomial Model

The unimodal model is a machine learning paradigm intended for supervised classification problems where the classes are ordered. Introduced in (Pinto da Costa et al., 2008), the main idea behind this model is that the random variable class associated with a given query should follow a unimodal distribution, so that the order relation between the classes is respected. In this context, the output of a classifier where the *a posteriori* class probabilities are estimated is obliged to be unimodal, *i.e.*, to have only one local maximum. There are different ways to impose unimodality and, in (Pinto da Costa et al., 2008), the authors suggested two approaches. In the parametric approach, a unimodal discrete distribution, like the binomial and Poisson's, is assumed and its parameters are estimated by the classifier. In the non-parametric approach, no distribution is assumed and the classifier is trained so that its output becomes unimodal. In all practical

experiments conducted by the authors, the parametric approach led to better results, in particular when the binomial distribution was considered. The superior performance achieved with this distribution was also justified in theoretical terms. For these reasons, our focus here is on the binomial model. Furthermore, since the classifiers chosen by us are artificial neural networks, support vector machines and decisions trees, we refer hereafter to binomial networks, binomial support vector machines and binomial tress, respectively. For the sake of conciseness, next, we only present a detailed description of the binomial networks applied to our problem.

As mentioned before, given information about a consumer, we are interested in predicting how much he/she is willing to spend in a bottle: less than EUR 2.99; between EUR 3 and 4.99; between EUR 5 and 9.99; EUR 10 or more. Representing these $K = 4$ bottle price classes by C_1, \dots, C_K , respectively, and the information given about the consumer by \mathbf{x} , Bayes' decision theory (Hastie et al., 2009) suggests classifying the case in the class maximizing the *a posteriori* probability $P(C_k|\mathbf{x})$. To that end, the *a posteriori* probabilities $P(C_1|\mathbf{x}), \dots, P(C_K|\mathbf{x})$ need to be estimated. In the binomial network, these probabilities are calculated from the binomial distribution $B(K-1, p)$. As this distribution takes values in the set $\{0, 1, \dots, K-1\}$, we take value 0 to represent class C_1 , 1 to C_2 , and so on, until $K-1$ to C_K . Now, since K is known, the only unknown parameter is the probability of success p . Hence, we consider a network architecture as in Figure 1 and train it to adjust all connection weights from layer 1 to layer 3. Note that the connections from layer 3 to layer 4 have a fixed weight equal to 1 and serve only to forward the value of p to the output layer of the network, where the probabilities from the binomial distribution are calculated. For a given query $\mathbf{x} = (x_1, \dots, x_J)$, the output of layer 3 will be a single numerical value in $[0, 1]$, denoted by $p_{\mathbf{x}}$. Then, the probabilities in layer 4 are calculated from the binomial distribution:

$$P(C_k|\mathbf{x}) = B_{k-1}(K-1, p_{\mathbf{x}}), k = 1, \dots, K, \quad (1)$$

where

$$B_{k-1}(K-1, p_{\mathbf{x}}) = \frac{(K-1)! p_{\mathbf{x}}^{k-1} (1-p_{\mathbf{x}})^{K-k}}{(k-1)! (K-k)!}. \quad (2)$$

When $p_{\mathbf{x}}$ is in $[0, \frac{1}{K}]$, the highest *a posteriori* probability is $P(C_1|\mathbf{x})$, and, therefore, the predicted bottle price class is C_1 . More generally, when $p_{\mathbf{x}}$ is in $[\frac{i-1}{K}, \frac{i}{K}]$, for some i in $\{1, \dots, K\}$, the highest *a posteriori* probability is $P(C_i|\mathbf{x})$, and, therefore, the predicted bottle price class is C_i . Hence, in order to train

the network on a training set $T = \{(\mathbf{x}_n, C_{\mathbf{x}_n})\}_{n=1}^N \subset \chi \times \{C_k\}_{k=1}^K$, where χ is the feature space, we replace C_k by the value of p corresponding to the midpoint of $[\frac{k-1}{K}, \frac{k}{K}]$, i.e., $p_k = \frac{k-0.5}{K}$, and apply a suitable optimization algorithm, like the Marquardt's method (Rao, 2019), to find connection weights that minimize the mean squared error

$$\frac{1}{N} \sum_{n=1}^N \left(p_{\mathbf{x}_n}^{target} - p_{\mathbf{x}_n}^{network}(\mathbf{w}) \right)^2, \quad (3)$$

where $p_{\mathbf{x}_n}^{target}$ is the value of p replacing $C_{\mathbf{x}_n}$ and $p_{\mathbf{x}_n}^{network}(\mathbf{w})$ is the output of layer 3 given the query \mathbf{x}_n and having the network the weights \mathbf{w} .

4.2 Modified Frank and Hall's Method

Frank and Hall's method was originally introduced in (Frank and Hall, 2001). Just like the unimodal model approach previously presented, the method is intended for supervised classification problems where the classes are ordered. As before, suppose that the $K = 4$ bottle price ordered classes are represented by C_1, \dots, C_K . Frank and Hall propose to use $K-1$ binary classifiers to address the K -class ordinal problem. In order to train the classifiers, such as artificial neural networks, support vector machines and decisions trees, $K-1$ data sets are derived from the original data set. The i -th classifier is trained to discriminate C_1, \dots, C_i from C_{i+1}, \dots, C_K . Given an unseen instance $\mathbf{x} = (x_1, \dots, x_J)$, i.e., information about a new consumer, the *a posteriori* probabilities $P(C_1|\mathbf{x}), \dots, P(C_K|\mathbf{x})$ of the original K classes can be estimated by combining the outputs of the $K-1$ binary classifiers for that instance. As noticed in (Cardoso and Pinto da Costa, 2007), the combination scheme suggested by Frank and Hall may lead to negative probabilities, but the problem can be overcome in the following manner: identifying the output p_i of the i -th classifier with the conditional probability $P(C_{\mathbf{x}} > C_i | C_{\mathbf{x}} > C_{i-1})$, the classes can be ranked according to the following formulas:

$$\begin{aligned} P(C_{\mathbf{x}} > C_1) &= p_1 \\ P(C_1|\mathbf{x}) &= 1 - p_1 \\ P(C_{\mathbf{x}} > C_j) &= p_j P(C_{\mathbf{x}} > C_{j-1}) \\ P(C_j|\mathbf{x}) &= (1 - p_j) P(C_{\mathbf{x}} > C_{j-1}), j = 2, \dots, K-1, \\ P(C_K|\mathbf{x}) &= P(C_{\mathbf{x}} > C_{K-1}). \end{aligned} \quad (4)$$

This is known as the modified Frank and Hall's method. Its implementation using networks is illustrated in Figure 2.

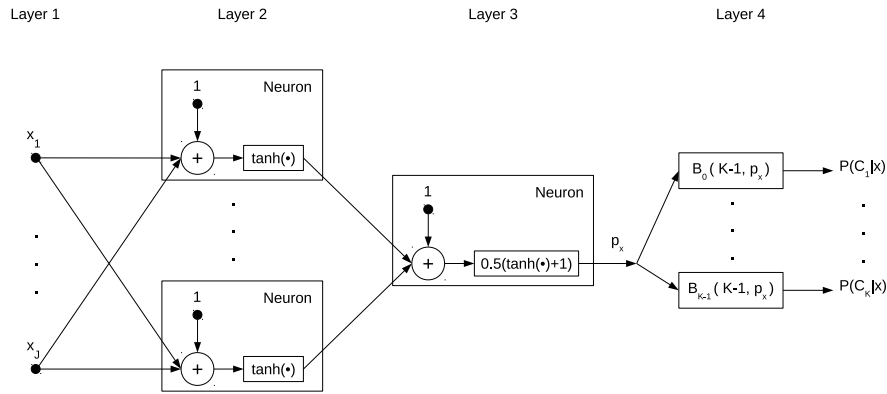


Figure 1: Binomial network.

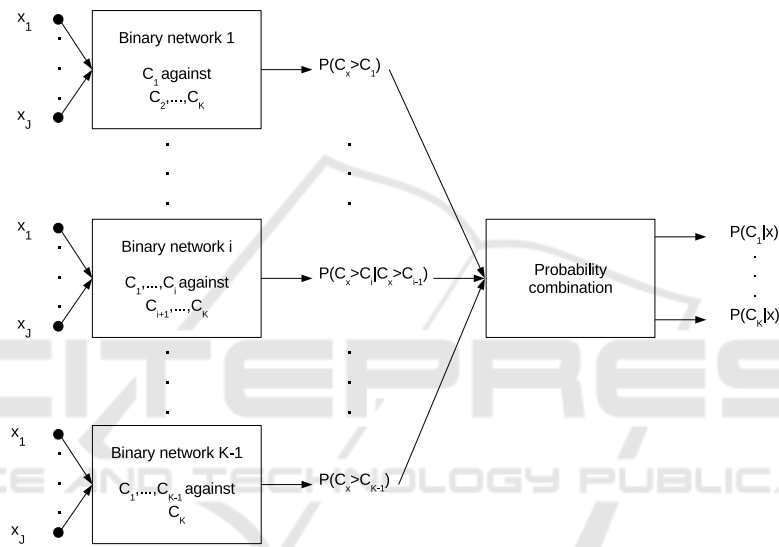


Figure 2: Implementation of the modified Frank and Hall's method using networks.

5 PERFORMANCE ASSESSMENT

The use of traditional metrics, like the overall accuracy, to assess the performance of a classifier in an imbalanced test set is not appropriate, because they tend to focus the model evaluation in the most frequent class(es) (Branco et al., 2016). In this section, we present suitable metrics for an imbalanced and multi-class problem like ours. For further details about the metrics and performance assessment in imbalanced domains, the reader should refer to (Branco et al., 2016) and references therein.

Suppose that there are n test instances. For the i -th test case, given the observed vector \mathbf{x}_i of the predictive attributes, a classifier makes a prediction $\hat{C}_{\mathbf{x}_i}$ of the true class $C_{\mathbf{x}_i}$. Let I be the indicator function that returns 1 if its argument is true and 0 otherwise. Then, the classifier has an overall accuracy or simply

accuracy given by

$$accuracy = \frac{\sum_{i=1}^n I(\hat{C}_{\mathbf{x}_i} = C_{\mathbf{x}_i})}{n}, \tag{5}$$

which corresponds to the proportion of cases that are correctly classified. As mentioned before, this is not an appropriate metric for an imbalanced test set. In an imbalanced and multi-class problem, if we focus on a single class C , then we can introduce the *recall* and *precision* for that class and the corresponding F_β score as

$$recall(C) = \frac{\sum_{i=1}^n I(C_{\mathbf{x}_i} = C) I(\hat{C}_{\mathbf{x}_i} = C)}{\sum_{i=1}^n I(C_{\mathbf{x}_i} = C)} \tag{6}$$

$$precision(C) = \frac{\sum_{i=1}^n I(C_{\mathbf{x}_i} = C) I(\hat{C}_{\mathbf{x}_i} = C)}{\sum_{i=1}^n I(\hat{C}_{\mathbf{x}_i} = C)} \tag{7}$$

$$F_\beta(C) = \frac{(1 + \beta^2) precision(C) recall(C)}{\beta^2 precision(C) + recall(C)}. \tag{8}$$

Hence, $recall(C)$ represents the proportion of cases from class C that are correctly classified and $precision(C)$ the proportion of cases predicted as being from class C that are correctly classified. Moreover, $F_\beta(C)$ is a combination of $recall(C)$ and $precision(C)$, where β is a parameter set by the user to adjust the relative importance of the former with respect to the latter. Usually, $\beta = 1$ and so the same importance is given to $recall(C)$ and $precision(C)$. Remark that $F_\beta(C)$ has a high value when both $recall(C)$ and $precision(C)$ are high. If there are K classes, C_1, \dots, C_K , one then averages $F_\beta(C_1), \dots, F_\beta(C_K)$ to obtain the so-called MF_β score:

$$MF_\beta = \frac{\sum_{k=1}^K F_\beta(C_k)}{K}. \quad (9)$$

This single scalar metric is considered suitable to compare the performance of different classifiers in an imbalanced test set. For this reason, we use it in the next section to analyze the results we obtained in our problem.

6 RESULTS

All computer experiments were carried out using Matlab R2021a with the Statistics and Machine Learning Toolbox. We fitted artificial neural networks (NNs), support vector machines (SVMs) and decision trees to perfectly balanced training data, obtained by applying the three re-sampling strategies previously described, namely random under sampling (RUS), random over-sampling with replacement (ROSWR) and the synthetic minority over-sampling technique for nominal predictive attributes (SMOTE-N). The models' hyperparameters, such as a regularization term strength in the case of NNs, the scale and type of kernel (Gaussian, linear or polynomial) in the case of SVMs and several parameters related to tree depth control in the case of decision trees, were chosen in order to obtain the best estimate of the prediction error, calculated by applying stratified 5-fold cross-validation to the training set (Hastie et al., 2009). In this way, we avoided underfitting and overfitting. This was done in the conventional approach to supervised classification, in the unimodal binomial paradigm (binomial model, for short) and in the modified Frank and Hall's method. The trained models were then applied to the test data.

Table 2 presents the results we obtained using the MF_1 score to assess the performance, in our test set, of different approaches to our problem of predicting how much a consumer is willing to pay for a bottle of wine. Remark that MF_1 is given by (9) with $\beta = 1$,

i.e., when we average F_β over all wine price classes to obtain MF_β , we give the same importance to the combination of $recall$ (6) and $precision$ (7) in F_β (8). Note that, the higher the value of these measures, the better. In Table 2, if we consider the results we obtained in (Alonso and Candeias, 2022) with the original imbalanced training data set, it can be seen the best one, $MF_1 = 0.4382$, was achieved by a NN in the modified Frank and Hall's method (best imbalanced classifier). If we look at the results we obtained in this paper by applying re-sampling strategies to balance the original training data set, it can be seen that, for each re-sampling strategy, the best result was also achieved by a NN in the modified Frank and Hall's method, with $MF_1 = 0.4836$ for RUS, $MF_1 = 0.4614$ for ROSWR and $MF_1 = 0.5528$ for SMOTE-N (best balanced classifier). Hence, we were able to increase the MF_1 score from 0.4382 in the imbalanced data approach to as much as 0.5528 when we applied re-sampling strategies.

Table 3 shows the F_1 score per class, in the test set, for the best imbalanced classifier and the best balanced one, and it can be seen that the application of the SMOTE-N re-sampling strategy led to an improvement of F_1 in the least represented classes in the test set (C_1, C_3 and C_4) and didn't decrease it significantly in the most represented one (C_2). Therefore, we achieved our goal of obtaining a more balanced classifier, *i.e.*, one with an improved ability to predict infrequent cases without seriously compromising the prediction of frequent ones.

The analysis we present next highlights the importance of a careful choice of the re-sampling strategy and the learning method in our imbalanced problem. From Table 2, remark that, in general, the use of RUS and ROSWR didn't improve the corresponding imbalanced data results when we considered the conventional and binomial learning methods; the only exception was in the conventional NN case, where the MF_1 score was slightly better with ROSWR. However, the use of RUS and ROSWR always improved the corresponding imbalanced data results when we considered the modified Frank and Hall's method, regardless of the type of classifier implemented; moreover, if we compare RUS with ROSWR, we can say that the former is preferable in most cases. Now, if we focus on the use of SMOTE-N, it is clear that, for all possibilities of learning method and classifier considered, it always led to results that are better than the ones obtained by applying RUS and ROSWR. Furthermore, if we compare it with the imbalance data approach, it can be seen that the only case where the results didn't improve was the one corresponding to the binomial NN. Finally, note that the best SMOTE-N results were

Table 2: Performance assessment in the test set using the MF_1 score.

| | | MF_1 | | | |
|---------------------------|------------|-----------------|----------------------|--------|---------|
| Learning method | Classifier | Imbalanced data | Re-sampling strategy | | |
| | | | RUS | ROSWR | SMOTE-N |
| Conventional | Tree | 0.3142 | 0.2592 | 0.2584 | 0.3680 |
| | SVM | 0.4355 | 0.4054 | 0.3411 | 0.4857 |
| | NN | 0.4089 | 0.3949 | 0.4152 | 0.4550 |
| Binomial | Tree | 0.3960 | 0.2107 | 0.3623 | 0.4236 |
| | SVM | 0.3807 | 0.2772 | 0.3479 | 0.4184 |
| | NN | 0.4141 | 0.2255 | 0.3527 | 0.3923 |
| Modified Frank and Hall's | Tree | 0.3464 | 0.3965 | 0.4387 | 0.4888 |
| | SVM | 0.3783 | 0.4756 | 0.3863 | 0.5386 |
| | NN | 0.4382 | 0.4836 | 0.4614 | 0.5528 |

Table 3: F_1 score per class, in the test set, for the best classifier fitted to the imbalanced training data (best imbalanced) and for the best classifier obtained by applying a re-sampling strategy to balance the training data, namely SMOTE-N (best balanced). The best classifiers are those exhibiting a higher MF_1 .

| Classifier | F_1 per class | | | |
|-----------------|-----------------|--------|--------|--------|
| | C_1 | C_2 | C_3 | C_4 |
| Best imbalanced | 0.6415 | 0.6667 | 0.4444 | 0.0000 |
| Best balanced | 0.7059 | 0.6269 | 0.5926 | 0.2857 |

always achieved when the modified Frank and Hall's method was set for learning algorithm. One difference between this method and the other ones lies in the fact that it is the only algorithm that, given an instance, combines the outputs of several classifiers in order to produce a final prediction of the true instance class; the other algorithms make the prediction based on only one classifier. We believe that this may be a reason for its success.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we presented a first approach to the issue of dealing with data imbalance in the multi-class ordinal classification problem of predicting how much a Portuguese consumer is willing to pay for a bottle of wine. More precisely, we applied several re-sampling strategies intended to balance the training data to which various predictive models were fit under different learning methods. In this context, we carried out a comparative study using performance measures adequate to the imbalance nature of the problem. We were able to obtain more balanced classifiers, *i.e.*, models with an improved ability to predict infrequent cases without seriously compromising the prediction of frequent ones. Furthermore, we concluded that the best balanced classifiers were the ones associated to the application of the SMOTE-N re-sampling strategy and the modified Frank and Hall's learning method.

Motivated by the good results of this method and the fact that it was the only one we applied where the outputs of several classifiers are combined in order to produce a final prediction of the true class, in the future, we plan to apply ensemble methods like bagging and boosting, where a set of individual learners are combined to create one learner with a better performance than the individual ones (see, for instance, (Galar et al., 2012; Tanha et al., 2020)). Moreover, we may apply a combination of under-sampling and over-sampling.

ACKNOWLEDGEMENTS

This work was partially supported by the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT – Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020.

REFERENCES

Alonso, H. and Candeias, T. (2022). Predicting how much a consumer is willing to pay for a bottle of wine: a preliminary study. In *Procedia Computer Science*, volume 204, pages 836–843.

Arafat, M. Y., Hoque, S., Xu, S., and Farid, D. M. (2019). Machine learning for mining imbalanced data. *IAENG*

- International Journal of Computer Science*, 46:332–348.
- Branco, P., Torgo, L., and Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49:1–50.
- Bruwer, J., Li, E., and Reid, M. (2002). Segmentation of the australian wine market using a wine-related lifestyle approach. *Journal of Wine Research*, 13:217–242.
- Cardoso, J. S. and Pinto da Costa, J. F. (2007). Learning to classify ordinal data: the data replication method. *Journal of Machine Learning Research*, 8:1393–1429.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.
- Cost, S. and Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10:57–78.
- Frank, E. and Hall, M. (2001). A simple approach to ordinal classification. In *Proceedings of the 12th European Conference on Machine Learning*, volume 1, pages 145–156.
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012). 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.
- Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2:42–47.
- Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems With Applications*, 73:220–239.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, USA, 2nd edition.
- Koksal, M. H. (2021). Segmentation of wine consumers based on level of involvement: a case of Lebanon. *British Food Journal*, 123:926–942.
- Kolyesnikova, N., Dodd, T. H., and Duhan, D. F. (2008). Consumer attitudes towards local wines in an emerging region: A segmentation approach. *International Journal of Wine Business Research*, 20:321–334.
- Kotler, P. and Keller, K. L. (2006). *Marketing management*. Prentice Hall, Upper Saddle River, USA, 12th edition.
- Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30:25–36.
- More, A. S. and Rana, D. P. (2021). Review of imbalanced data classification and approaches relating to real-time applications. In Rana, D. and Mehta, R., editors, *Data Preprocessing, Active Learning, and Cost Perceptive Approaches for Resolving Data Imbalance*, chapter 1, pages 1–22. IGI Global, Pennsylvania, United States.
- Payini, V., Bolar, K., Mallya, J., and Kamath, V. (2022). Modeling hedonic motive-based segments of wine festival visitors using decision tree approach. *International Journal of Wine Business Research*, 34:19–36.
- Pinto da Costa, F., J., Alonso, H., and Cardoso, J. S. (2008). The unimodal model for the classification of ordinal data. *Neural Networks*, 21:78–91.
- Pinto da Costa, F., J., Alonso, H., and Cardoso, J. S. (2014). Corrigendum to ‘The unimodal model for the classification of ordinal data’ [Neural Netw. 21 (2008) 78–79]. *Neural Networks*, 59:73–75.
- Rao, S. S. (2019). *Engineering Optimization: Theory and Practice*. John Wiley & Sons, Inc, New Jersey, USA, 5th edition.
- Rouzet, E. and Seguin, G. (2004). *Il marketing del vino. Il mercato. Le strategie commerciali. La distribuzione*. Il Sole 24 ORE Edagricole, Bologna, Italia.
- Stanfill, C. and Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM*, 29:1213–1228.
- Sun, Y., Wong, A. K. C., and Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23:687–719.
- Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., and Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, 7:1–47.
- Thach, E. C. and Olsen, J. E. (2005). The search for new wine consumers: Marketing focus on consumer lifestyle or lifecycle? *International Journal of Wine Marketing*, 16:44–57.
- Thach, E. C. and Olsen, J. E. (2006). Market segment analysis to target young adult wine drinkers. *Agribusiness*, 22:307–322.