University of Alberta, CMPUT 551 Fall 2017

Mini-project Report: Examining Sampling Methods Performance for Imbalanced Data

Humza Haider, Student ID: 1535075
December 8, 2017

Introduction and Data Overview

Imbalanced classification can be extremely problematic for machine learning algorithms as it forces the algorithm to learn a model which will likely bias classifications toward the majority class. A typical example of the class imbalance problem is seen in finance when trying to determine which transactions out of a sample are real and which ones are fraudulent. A plausible case could be a random sample containing thousands of real transactions and a few hundred or less fraudulent transactions. When models are learned using data such as these, they often predict all observations to be of the majority class. This paper examined the effects of different sampling techniques and these techniques' impact on the predictive power of the learned model.

The data analyzed came from the Abalone Data Set from the UCI Machine Learning Repository[1]. An abalone is a type of sea snail that is distributed across the world in the coastal waters of every continent. The age of an abalone can be found by cutting the shell open, staining the shell, and counting the rings via microscope. The process of obtaining the age of an abalone is time consuming and serves as a candidate which could benefit from automated prediction. The dataset contained 4177 observations and 9 features, including the age. The features with summary statistics are given in Table 1.

Variable	Description	Mean/Proportion	Standard deviation
Sex	Male	0.366	NA
	Female	0.313	NA
	Infant	0.321	NA
Length	Longest shell measurement - mm	0.524	0.120
Diameter	Distance perpendicular to length - mm	0.408	0.099
Height	Height with meat in shell - grams	0.140	0.042
Whole Weight	Weight of the entire abalone - grams	0.829	0.490
Shucked Weight	Weight of the meat - grams	0.359	0.222
Viscera Weight	Gut weight (after bleeding) - grams	0.181	0.110
Shell Weight	Weight after being dried - grams	0.239	0.139
Age - years	Given by a positive integer.	9.934	2.322

Table 1: Summary statistics of the 9 variables made available in the dataset. All features with the exception of Sex are numerical. All numerical variables are given on a continuous scale with the exception of age which is given by strictly positive integers.

Given that the age of an abalone takes values which are strictly positive integers it seems intuitive to use Poisson regression to model the data. However, this problem can also be modeled as a classification problem by considering a specific age to be the positive class and all other ages to belong to the negative class. This data was chosen specifically for this reason. The primary research goal of this paper was to take a highly imbalanced classification dataset and attempt

[1]https://archive.ics.uci.edu/ml/datasets/abalone

to find a sampling method which could perform better than a model with no sampling. To form the imbalanced dataset, the age of 11 was chosen to be the positive class and all other ages were combined to form the negative class. In total, 487 and 3690 observations were labeled as the positive and negative class, respectively. Given these numbers, the resulting class imbalance was 11.66% positive observations and 88.34% negative observations.

Methods

This section overviews the different sampling procedures applied to the data. Further, the scoring metric used to compare the models is chosen and justified. Following this justification, an overview of the model hyperparameters and experimental design is provided.

Sampling Procedures

When approaching the class imbalance problem, this paper focused on sampling procedures as opposed to different learning algorithms. The intuition of the sampling procedure is to sample the training data in such a way that the model is trained from a class balanced data set. The first sampling procedure applied is meant to over sample the minority class - referred from here on as *up-sampling*. Given the abalone data, up-sampling randomly samples, with replacement, the 487 positive observations until the same number of positive and negative classes is achieved, *i.e.* until 3690 positive observations were sampled. Note that the original 487 are included and then added to through random sampling.

The next sampling approach was to under sample the majority class, referred to as *down-sampling*. Given the abalone data, down-sampling randomly selected 487 of the negative observations and used these along with the 487 positive observations to learn the model. In both up-sampling and down-sampling the final training dataset contains the same number of positive and negative observations.

The last sampling method is known as the Synthetic Minority Over-sampling Technique (SMOTE) [2]. The SMOTE method creates synthetic observations by calculating the positive observation k-nearest neighbors of a positive observation, randomly selecting one of the neighbors and taking the difference of the feature vector of the original observation and it's chosen neighbor. This difference is then multiplied by a random number between 0 and 1, exclusive, and then added back to the original observation. For this analysis, k was tested for values from the set $\{5, 10, 15, 20, 25\}$. The limit of 25 was chosen as this represents roughly the nearest 5% of neighbors to the original positive observation, thus limiting large differences between synthetic and real positive observations.

The SMOTE method takes as an argument the percentage to up-sample and creates the proper number of synthetic observations accordingly. For example, if 200% was the percentage to up-sample, the total positive observations would contain the original 487 positive observations and 487 synthetically generated positive observations. In addition, the SMOTE method can be paired with down-sampling. If the chosen percentage to down-sample the majority class was 100%, there would exist the same number of positive and negative observations. To further clarify the arguments of the SMOTE method consider hyperparameters such that SMOTE is applied with an up-sample of 200% , a down-sample of 300%, and $k = 10$ neighbors. The final data set would have $2*487 = 974$ positive observations and $3*487 = 1461$ negative observations. These extra 487 positive cases would be generated synthetically as explained above using $k = 10$ nearest neighbors. For this analysis, the down and up-sample values belonged to 100%, 200%, 300%, 400%, 500%, 600%, and 700%. The SMOTE sampling percentages are given a cutoff of 700% since 700% down-sampling of the negatives represents 3409 negative cases of the original

3690. To use 800% or above would implicate the creation of synthetic negative observations which is counter intuitive as the data already contains a large number of negative observations.

It is important to note that this sampling is often misused when paired with cross-validation. In order to get a proper estimate of prediction error one must create the cross-validation folds and then apply the sampling procedure. This does not have large impacts for down-sampling but can largely bias the up-sampling estimates since a fair number of the same observations will be used for training and testing if sampling is applied prior to cross-validation folds. For a more rigorous understanding of this effect, please refer to Lusa and Blagus (2015) [3].

Scoring Metric

Typically, accuracy serves as a standard scoring metric, however, given the class imbalance in the abalone data, accuracy gives misleading performance of the models predictive ability. Consider a model which only predicts negative classes. The accuracy of this model applied to the abalone data will attain 88.34% without having learned anything from the data. Further, estimates of precision and recall require a specific cut-off value, or threshold, in order to be computed, whereas an ideal choice for the threshold may vary across different models. To rank different models, a metric which considers all values of this threshold must be used.

A common metric with this property is the Area Under the Receiver Operating Characteristic Curve (AUROCC). However, the AUROCC metric relies on the recall and the false positive rate to be computed. Since the false positive rate is used instead of precision, the AUROCC is invariant under class imbalances as the false positive rate does not significantly drop when the number of real negatives is very large. For a further discussion of this please see the footnote provided[2].

Due to the caveats in the previously mentioned metrics, the area under the precision recall curve (AUPRC) was settled upon as the scoring metric. The AUPRC is very similar to the AUROCC but uses precision instead of the false positive rate. Since precision is computed using true and false positives, precision is therefore sensitive to models which tend to classify all observations as negatives with high probability. [1] By making this exchange, the AUPRC takes into account the class imbalance of the data and serves as a proper comparison metric between models. Examples of the precision recall curve applied to the abalone data can be found in Figure 1.

Model Selection and Cross-Validation

For each sampling method, logistic regression with L2 regularization was applied with the regularization parameter, $\lambda \in \{0.0, 0.1, 0.01, 0.001, 0.0001\}$. These values were chosen to represent varying degrees of bias introduced into the model to limit the variance of model parameters. Since multiple estimates of AUPRC are required to determine significant differences between sampling methods, Monte-Carlo cross-validation was applied. This form of cross-validation randomly selects a portion of the observations to be the training set and assigns the rest of the observations to be the test set. As per usual, the model is then trained on the training set and analyzed on the test set. However, this process will be repeated for a specified number of times. For these simulations, Monte-Carlo cross-validation was implemented with 10-folds, i.e. 10 randomly sampled training and test sets were formed. Each training and test set followed the proportions 70% training data and 30% test data. In order to select hyperparameters, standard 10-fold cross validation was used on each training set and the best performing hyperparameters, in terms of AUPRC, were selected for use on the test set.

[2] https://classeval.wordpress.com/simulation-analysis/roc-and-precision-recall-with-imbalanced-datasets/

The Analysis of One Way Variance (ANOVA) test is a standard statistical test used to analyze differences among group means. The null hypothesis claims all group means are equal, whereas the alternative hypothesis is that not all means are equal. For this paper, the ANOVA test is used to analyze if there are significant differences between the mean AUPRC of the 3 different sampling methods and the model with no sampling method. The 'best' performing model would be the model which has the highest AUPRC among the methods considered. Given a result which showed a tie between models, the most computationally efficient model would be selected as the 'best'. All analyses were conducted in R version 3.4.1.

Results

Empirical results of the four different models are made available in Table 2. Note that for a threshold value of 0.500, when no sampling method is applied to the data, logistic regression will classify every observation as a negative class, resulting in a recall of 0.000 and an incalculable precision, due to the lack of true and false positives. Note that while the different sampling techniques are able to remedy this problem, the accuracy drops staggeringly from 0.883 to 0.516-0.597. Further, even with the very poor value for precision for the model with no sampling, all models perform nearly identically in terms of the AUROCC.

For nearly every fold, the value of λ was 0 for all algorithms, suggesting regularization was not influential for logistic regression on this dataset, regardless of sampling technique. Interestingly, for SMOTE sampling, the optimal value for k was 25, the maximum value, for every fold of Monte-Carlo cross-validation. In future tests, allowing the value of k to take greater values could impact the performance of the SMOTE method on this data. Further, the SMOTE method produced up and down-sampling values of 300 and 600 for over half the folds in the Monte-Carlo cross-validation. These values indicate 1461 positive observations (974 synthetically generated, 487 original) and 2322 negative cases. The intuition of why these were common choices of as optimal hyperparameters is unclear.

An example of the precision recall curve is given for each sampling technique in Figure 1. This figure was generated by taking the model performance on the test set of one of the Monte-Carlo cross-validation folds. The precision recall curves are nearly identical across the sampling techniques, only differing in the threshold values used across the model. Maximum precision was attained at a threshold of roughly 0.35, 0.70, 0.85, and 0.40 for no sampling, up-sampling, down-sampling and SMOTE, respectively.

Sampling	AUPRC	AUROCC	Accuracy	Recall	Precision
None	0.234 ± 0.030	0.706 ± 0.021	0.883 ± 0.000	0.000 ± 0.000	NA
Up	0.229 ± 0.026	0.705 ± 0.021	0.596 ± 0.003	0.684 ± 0.039	0.178 ± 0.005
Down	0.224 ± 0.032	0.705 ± 0.022	0.597 ± 0.003	0.683 ± 0.037	0.178 ± 0.005
SMOTE	0.225 ± 0.027	0.704 ± 0.022	0.516 ± 0.011	0.772 ± 0.037	0.164 ± 0.002

Table 2: Means ± standard deviation for results pertaining to logistic regression and the sampling methods employed. Precision, recall, and Accuracy are given for a threshold value of 0.50.

The ANOVA test comparing the mean AUPRC scores across the different sampling methods yielded a p-value of 0.881, suggesting that none of the sampling methods had a statistically significant effect on the AUPRC. Given this finding, the optimal model choice of the four examined would be the logistic regression with no sampling applied as this model is the computationally simplest.

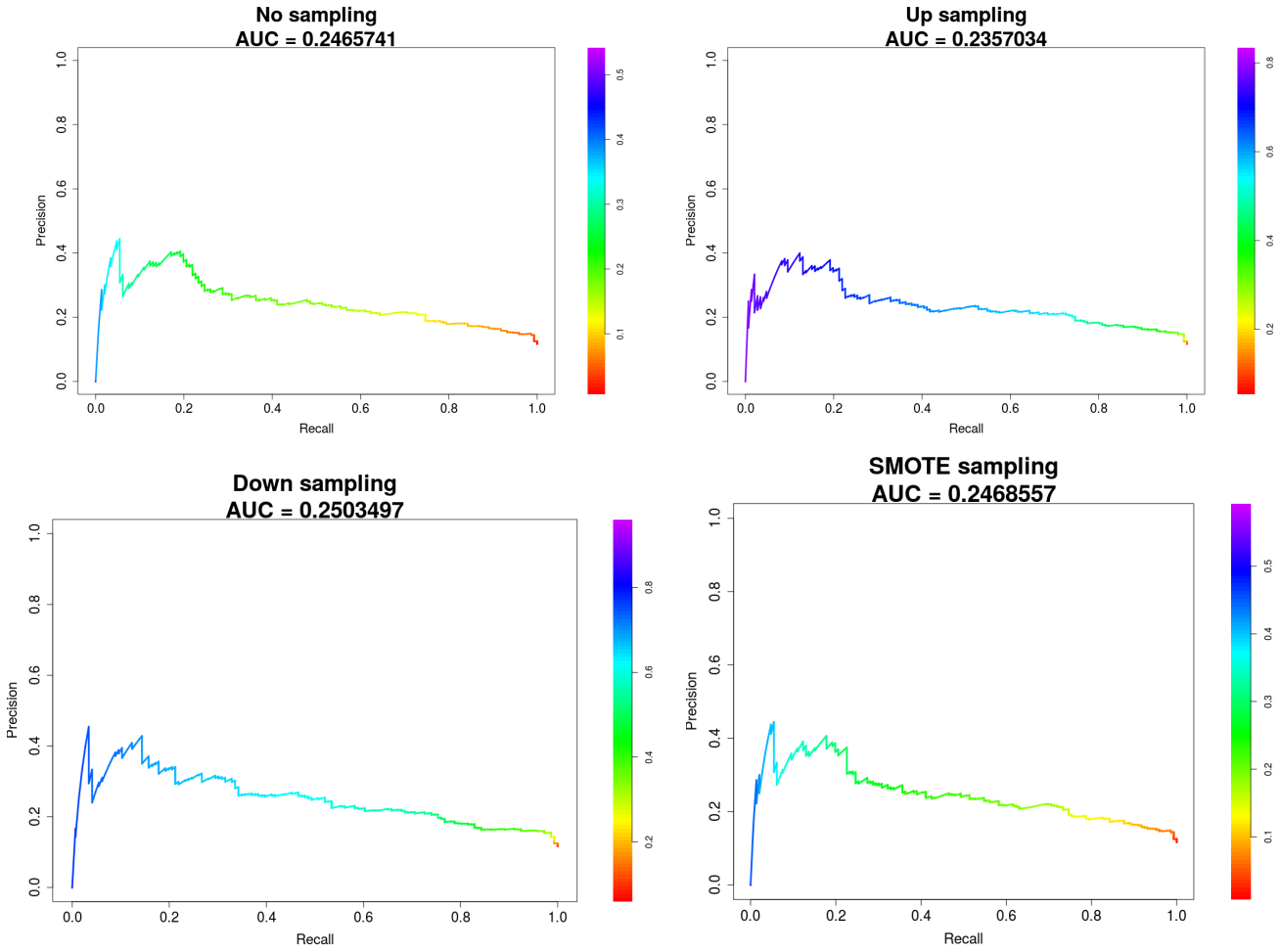

Figure 1: Sample PR curve from one training/test sample.

Conclusion

Three different methods of sampling were employed to predict a classification problem with imbalanced data, specifically with 11.66% positive observations and 88.34% negative observations. Overall, none of the sampling methods were able to achieve statistically significant differences in their AUPRC scores, suggesting that the sampling procedures had no significant effect on the predictive power of logistic regression, in terms of AUPRC, for this data. Given these findings, the simplest model, with no sampling procedure, was chosen as the final model which produced an AUPRC score of 0.234 ± 0.030. Future work using this data should allow higher values of k for the SMOTE method in order to see if this impacts the model performance. Further, work could be done to test how the sampling procedures perform for learning models such as neural networks, naive Bayes, support vector machines and other classification models.

References

[1] Kendrick Boyd, Kevin H Eng, and C David Page. "Area under the precision-recall curve: Point estimates and confidence intervals". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2013, pp. 451–466.

[2] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[3] Lara Lusa et al. "Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models". In: *BMC bioinformatics* 16.1 (2015), p. 363.