Supplementary material to the paper "Is augmentation effective in improving prediction in imbalanced datasets?"

Gabriel O. Assunção
Department of Statistics, Universidade Federal de Minas Gerais,
Belo Horizonte, Brazil
and
Rafael Izbicki
Department of Statistics, Universidade Federal de São Carlos,
São Carlos, Brazil
and
Marcos O. Prates
Department of Statistics, Federal de Minas Gerais,
Belo Horizonte, Brazil

September 16, 2024

S.1 Proofs

In this section, we will present the proof of Theorems and Corollary of Section 2.

Proof of Theorem 1. We assume x is discrete, although the proof is always valid.

$$\mathbb{P}_{\mathcal{T}}(X = x|Y = k) = \mathbb{P}_{\mathcal{T}'}(X = x|Y = k).$$

It follows that

$$
\begin{aligned}
\mathbb{P}_a(\{Y=k\}\cap\{X=x\}) = &\frac{n}{n+m}\mathbb{P}_{\mathcal{T}}(\{Y=k\}\cap\{X=x\})+ \\
&\frac{m}{n+m}\mathbb{P}_{\mathcal{T}'}(\{Y=k\}\cap\{X=x\}) \\
&= \frac{n}{n+m}\mathbb{P}_{\mathcal{T}}(X=x|Y=k)\mathbb{P}_{\mathcal{T}}(Y=k)+ \\
&\frac{m}{n+m}\mathbb{P}_{\mathcal{T}'}(X=x|Y=k)\mathbb{P}_{\mathcal{T}'}(Y=k) \\
&= \left(\frac{n}{n+m}\mathbb{P}(Y=k)+\frac{m}{n+m}w_k\right)\mathbb{P}(X=x|Y=k) \\
&= \mathbb{P}_a(Y=k)\frac{\mathbb{P}(Y=k|x)\mathbb{P}(x)}{\mathbb{P}(Y=k)}.
\end{aligned}
$$

By replacing this expression in

$$
\mathbb{P}_a(Y=k|x) = \frac{\mathbb{P}_a(\{Y=k\}\cap\{X=x\})}{\sum_k \mathbb{P}_a(\{Y=k\}\cap\{X=x\})},
$$

we conclude that

$$
\begin{aligned}
\mathbb{P}_a(Y=k|x) &= \frac{\mathbb{P}_a(\{Y=k\}\cap\{X=x\})}{\sum_k \mathbb{P}_a(\{Y=k\}\cap\{X=x\})} \\
&= \frac{\frac{\mathbb{P}_a(Y=k)}{\mathbb{P}(Y=k)}\mathbb{P}(Y=k|x)\mathbb{P}(x)}{\sum_k \frac{\mathbb{P}_a(Y=k)}{\mathbb{P}(Y=k)}\mathbb{P}(Y=k|x)\mathbb{P}(x)} \\
&= \frac{\frac{\mathbb{P}_a(Y=k)}{\mathbb{P}(Y=k)}\mathbb{P}(Y=k|x)}{\sum_k \frac{\mathbb{P}_a(Y=k)}{\mathbb{P}(Y=k)}\mathbb{P}(Y=k|x)}.
\end{aligned}
$$

□

Proof. Theorem 2

By putting together Equations (1) and (2), we conclude that

$$
g^*(\mathbf{x}) := \arg\min_{j\in\mathcal{Y}} \sum_{k\in\mathcal{Y}} L'(k,j)\frac{\mathbb{P}(Y=k|\mathbf{x})}{h(\mathbf{x})},
$$

where $h(\mathbf{x}) = \sum_{k=1}^{K} \frac{\mathbb{P}_{\mathrm{a}}(Y=k)}{\mathbb{P}(Y=k)}\mathbb{P}(Y=k|\mathbf{x})$. The conclusion follows from the fact that h is constant in j and k. □

Corollary 1. In the binary case, with an optimal classifier g^* according to the induced probability and 0-1 loss. In the case where $\mathbb{P}_a(Y=1)=0.5$, the loss function that gives the same classifier on the original probability is

$$
L'(1,0) = \frac{\mathbb{P}_{\mathrm{a}}(Y=1)}{\mathbb{P}(Y=1)} = \frac{1}{2\mathbb{P}(Y=1)}
$$

and

$$L'(0,1) = \frac{\mathbb{P}_{\text{a}}(Y=0)}{\mathbb{P}(Y=0)} = \frac{1}{2\mathbb{P}(Y=0)}.$$

Based on Propostion 2.1 (Ripley, 2007), it follows that the decision rule corresponds is:

$$\begin{aligned}
\frac{L'(0,1)}{L'(0,1)+L'(1,0)} &= \frac{\frac{1}{2\mathbb{P}(Y=0)}}{\frac{1}{2\mathbb{P}(Y=0)}+\frac{1}{2\mathbb{P}(Y=1)}} \\
&= \frac{\frac{1}{2\mathbb{P}(Y=0)}}{\frac{\mathbb{P}(Y=1)+\mathbb{P}(Y=0)}{2\mathbb{P}(Y=0)\mathbb{P}(Y=1)}} \\
&= \frac{2\mathbb{P}(Y=0)\mathbb{P}(Y=1)}{2\mathbb{P}(Y=0)} \\
&= \mathbb{P}(Y=1)
\end{aligned}$$

To demonstrate that this classifier is the one that maximizes the balanced accuracy, we use the relation presented in Izbicki and dos Santos (2020, pg 170) that the classifier

$$g^*(x) = \mathbb{I}\left(\mathbb{P}(Y=1|\mathbf{X}) > \frac{l_1}{l_1+l_0}\right)$$

is the one that minimize the risk function

$$R(g^*) = \mathbb{E}[l_1\mathbb{I}(Y=0, g^*(\mathbf{X})=1) + l_0\mathbb{I}(Y=1, g^*(\mathbf{X})=0)].$$

Replacing $l_1 = L'(0,1) = \frac{1}{\mathbb{P}(Y=0)}$ e $l_0 = L'(1,0) = \frac{1}{\mathbb{P}(Y=1)}$ we have the same classifier than Equation (5). We need to show that when minimizing this risk we obtain the best balanced accuracy:

$$\begin{aligned}
R(g^*) &= \mathbb{E}[L'(0,1)\mathbb{I}(Y=0, g^*(\mathbf{X})=1) + L'(1,0)\mathbb{I}(Y=1, g^*(\mathbf{X})=0)] \\
&= L'(0,1)\mathbb{P}(Y=0, g^*(\mathbf{X})=1) + L'(1,0)\mathbb{P}(Y=1, g^*(\mathbf{X})=0) \\
&= \frac{\mathbb{P}(Y=0, g^*(\mathbf{X})=1)}{\mathbb{P}(Y=0)} + \frac{\mathbb{P}(Y=1, g^*(\mathbf{X})=0)}{\mathbb{P}(Y=1)} \\
&= \mathbb{P}(g^*(\mathbf{X})=1|Y=0) + \mathbb{P}(g^*(\mathbf{X})=0|Y=1).
\end{aligned}$$

Therefore, when we minimize this risk function we maximize

$$\mathbb{P}(g^*(\mathbf{X})=1|Y=1) + \mathbb{P}(g^*(\mathbf{X})=0|Y=0),$$

that by definition this is the balanced accuracy.

□

S.2 Hypothesis test

In this section, we present the hypothesis text adopted in the main manuscript to check the difference in the performance of the original and augmented databases.

Fix an augmentation method and a dataset, and let $X_{i,j}$ be the percentage gain on the j-th augmented dataset for the i-th sample of the original data, $i = 1, \ldots, 50$ and $j = 1, \ldots, 40$. In order to take the dependency between measurements obtained on the same sample of the original data, we assume that $X_{i,j} \sim N(M_i, \sigma_R^2)$ are independent random variables given $M_1, \ldots, M_5$, that $M_1, \ldots, M_5 \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, and that $\mu, \sigma^2, \sigma_R^2$ are fixed parameters.

Our goal is to test the null hypothesis $H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$. First, we compute the test statistic $T = \frac{1}{2000} \sum_{i,j} X_{i,j}$, the average percentage gain. Then, we compute p-values based on T via a parametric bootstrap. This is done by first estimating σ^2 and σ_R^2 using their maximum likelihood estimates, $\widehat{\sigma}^2, \widehat{\sigma}_R^2$. Then, we sample the test statistic from the null by sample data with the same structure as $X_{i,j}$'s at the point $(\mu, \sigma^2, \sigma_R^2) = (0, \widehat{\sigma}^2, \widehat{\sigma}_R^2)$ and computing the test statistic for each sampled dataset, $T^{(1)}, \ldots, T^{(B)}$. The p-value is simply

$$\frac{1}{B} \sum_{b=1}^{B} I\left(|T^{(b)}| \geq |T|\right).$$

We take $B = 1000$.

S.3 Logistic Analysis

In this section, we present the results of training with a logistic model and compare the model with the same comparative criteria presented in Section 3. Figure S.1 displays the percentage gain in balanced accuracy for the logistic regression. The results are similar as the obtained with the Random Forest.

Figure S.2 shows the gain on the AUC and Brier Score. The results indicate that the estimation is better with the non-augmented method.

S.4 ROC curve analysis

This section presents the ROC curve analysis shown in Section 3 for the other datasets. The top row shows the settings with greater gains in the AUC, while the bottom row the cases with the lowest gain. Overall the conclusions are similar to the ones discussed in the main manuscript.

S.5 Other metrics analysis

We evaluate the percentage gain over other metrics. In this section, we displayed the results for the F1-score, Accuracy, Sensitivity, and Specificity. The outcomes for F1-score

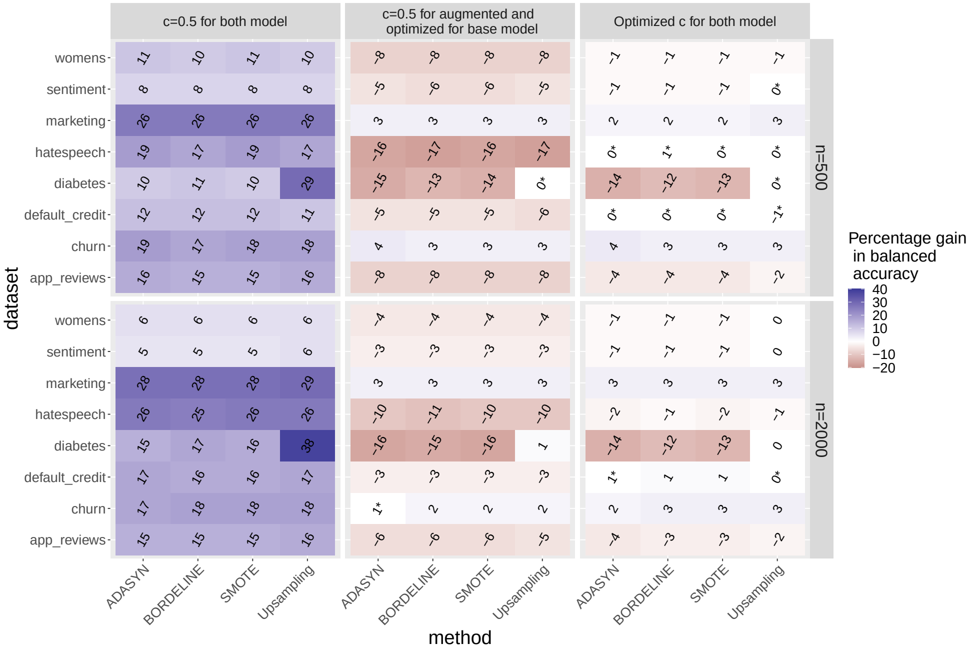

Figure S.1: Heatmap of the mean percentage gain in balanced accuracy when comparing the augmented methods with the non-augmented model for classification rules. Positive values indicate superior performance by the augmented method. Non-significant gains are marked with asterisks and displayed in white. Our findings indicate that with the logistic model optimizing the threshold eliminates the need for augmentation.

and sensitivity exhibited a resemblance to those obtained for balanced accuracy. Although the default threshold produced a perceived enhancement, optimization of the threshold led to similar outcomes. In contrast, while utilizing the default threshold did not lead to improvements with augmented methods for specificity and accuracy, the optimized threshold showed an improvement in the augmented model, particularly in terms of specificity.

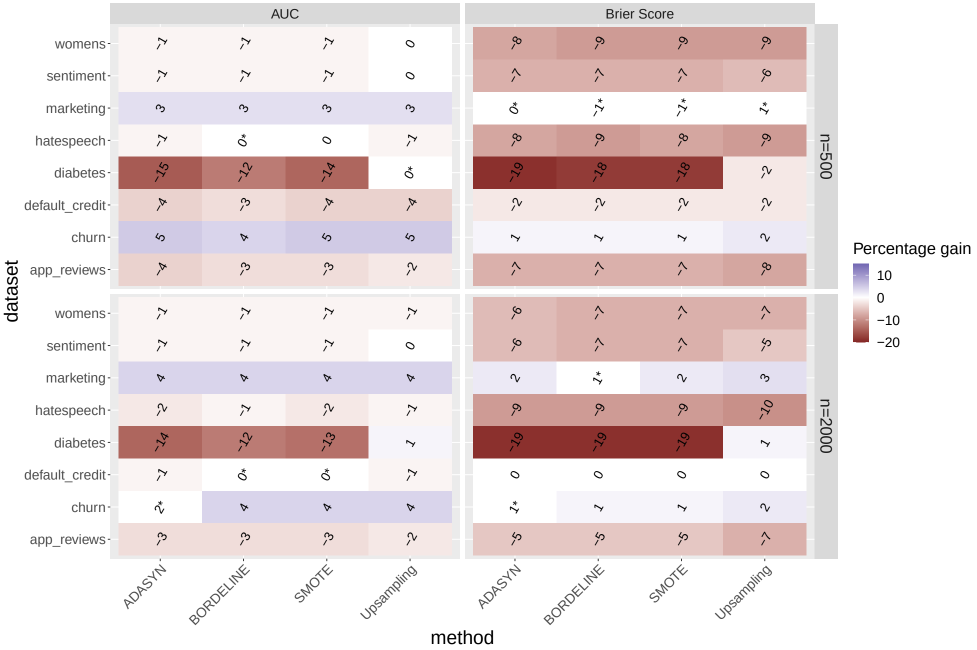

Figure S.2: Heatmap of the average percentage improvement in the AUC (left column) and Brier Score (right column) when comparing the augmented methods with the non-augmented ones. Positive values indicate superior precision in estimating $\mathbb{P}(Y = 1|\mathbf{x})$ using the augmented method. Non-significant gains are marked with asterisks and displayed in white. The results indicate that data augmentation never improves $\mathbb{P}(Y = 1|\mathbf{x})$ estimates.

References

Izbicki, R. and dos Santos, T. M. (2020). *Aprendizado de Máquina: Uma Abordagem Estatística*.

Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge University Press.

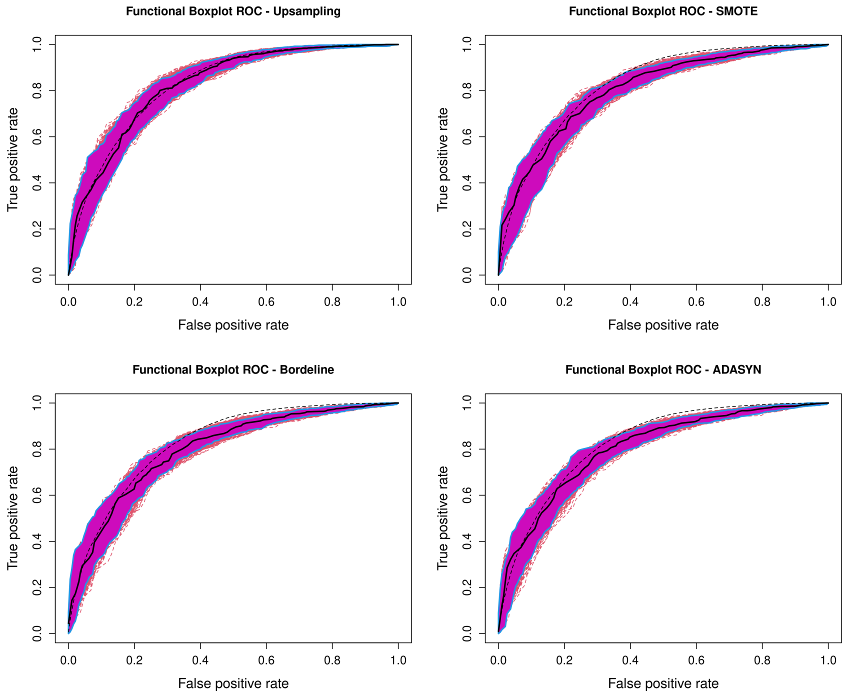

Figure S.3: Functional BoxPlot of the Churn dataset on the train size of 2000

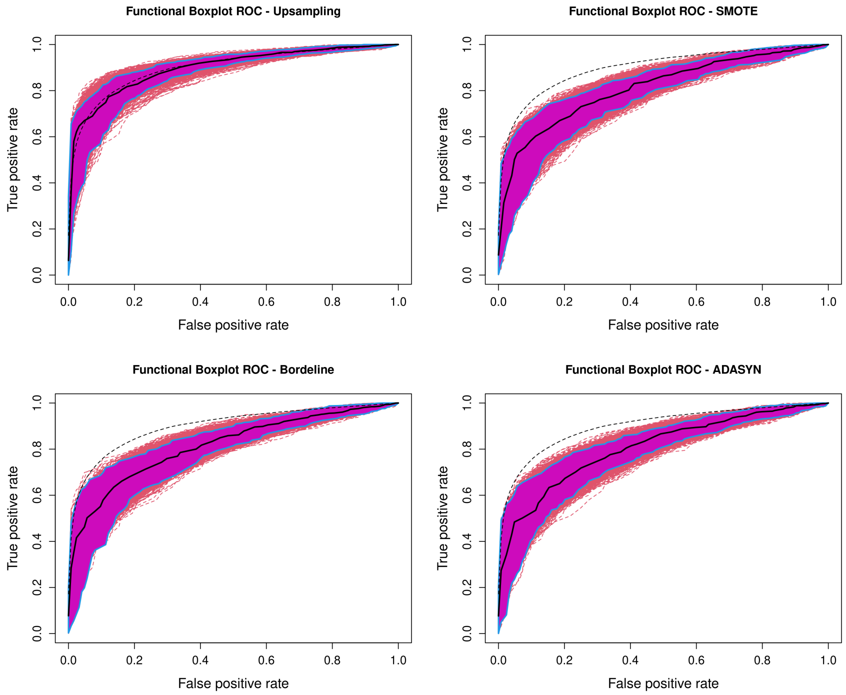

Figure S.4: Functional BoxPlot of the Marketing dataset on the train size of 500

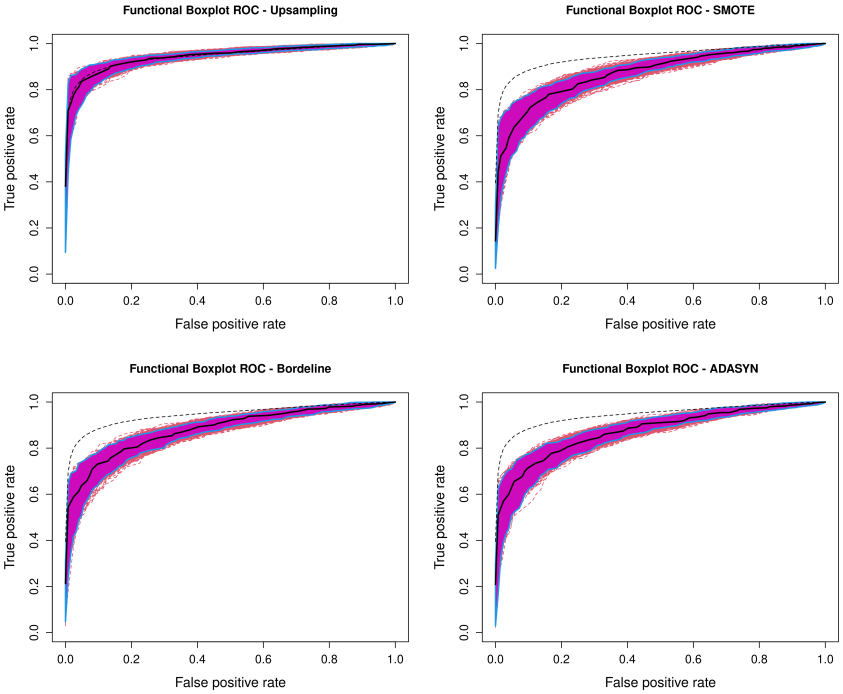

Figure S.5: Functional BoxPlot of the Marketing dataset on the train size of 2000

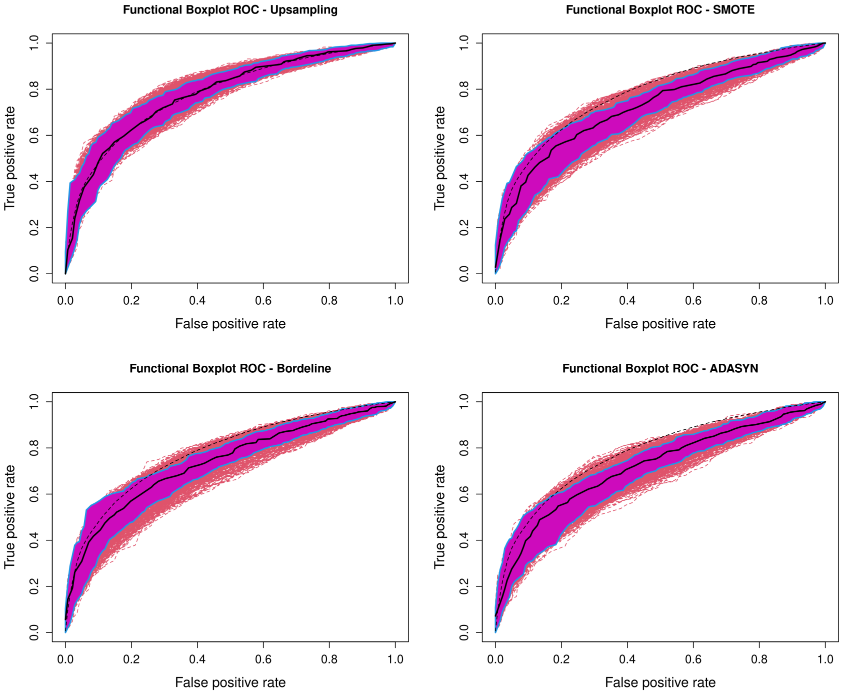

Figure S.6: Functional BoxPlot of the Diabetes dataset on the train size of 500

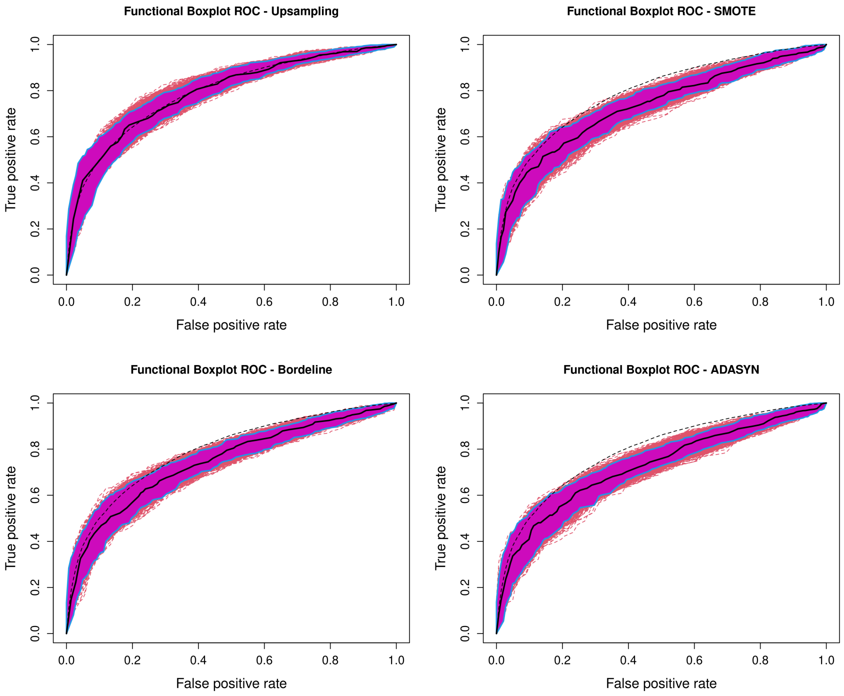

Figure S.7: Functional BoxPlot of the Diabetes dataset on the train size of 2000

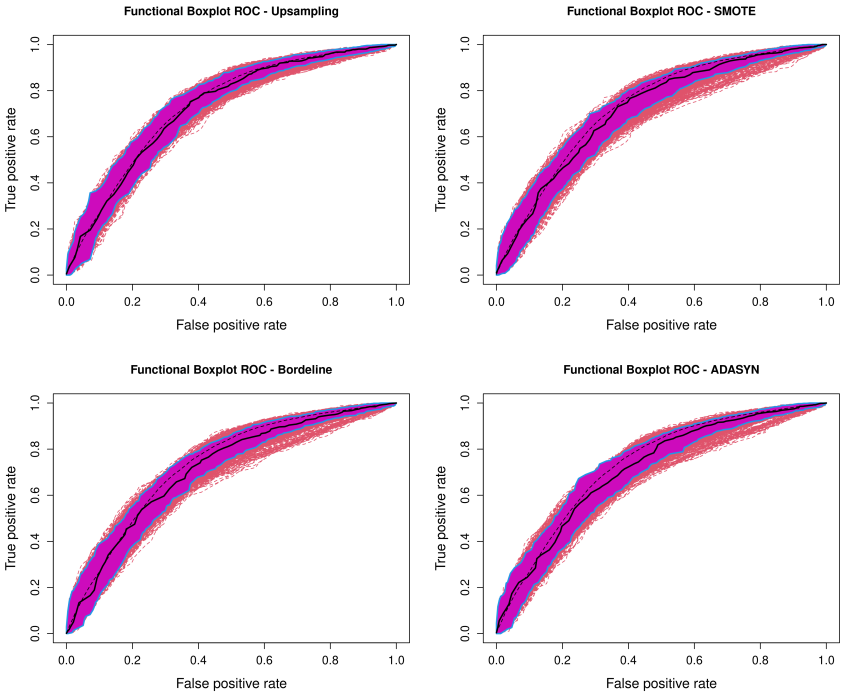

Figure S.8: Functional BoxPlot of the Default Credit dataset on the train size of 500

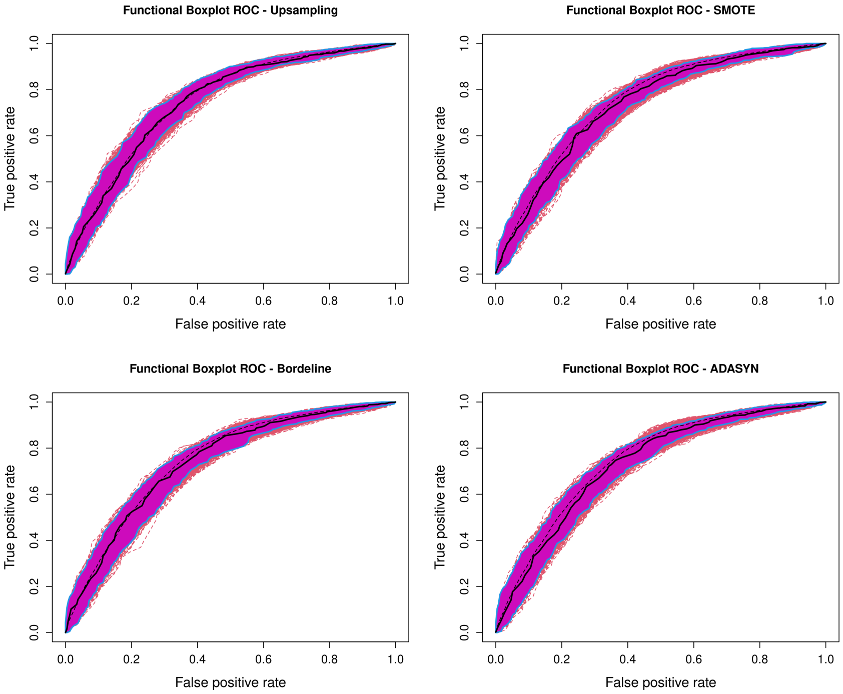

Figure S.9: Functional BoxPlot of the Default Credit dataset on the train size of 2000

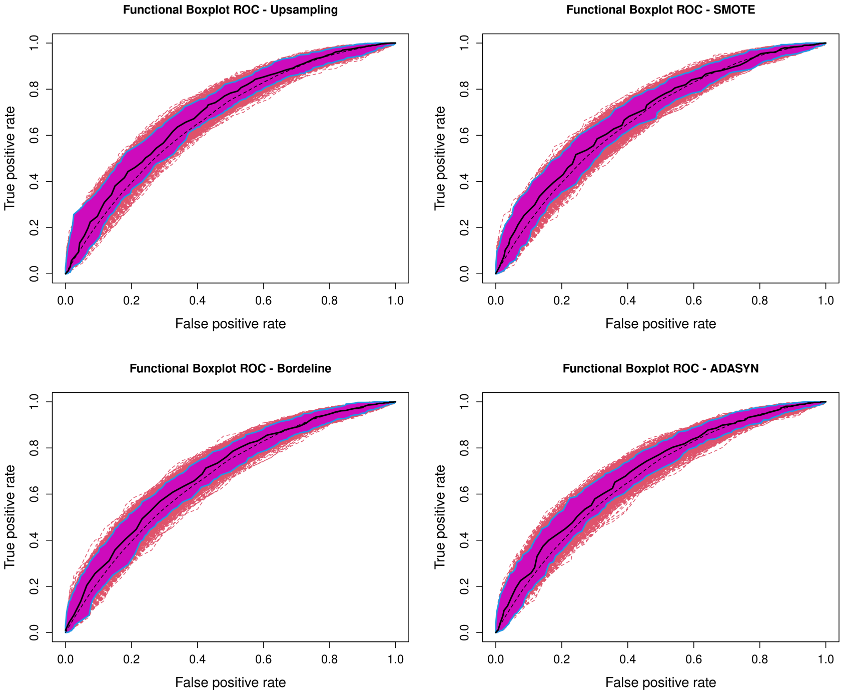

Figure S.10: Functional BoxPlot of the Sentiment Twitter dataset on the train size of 500

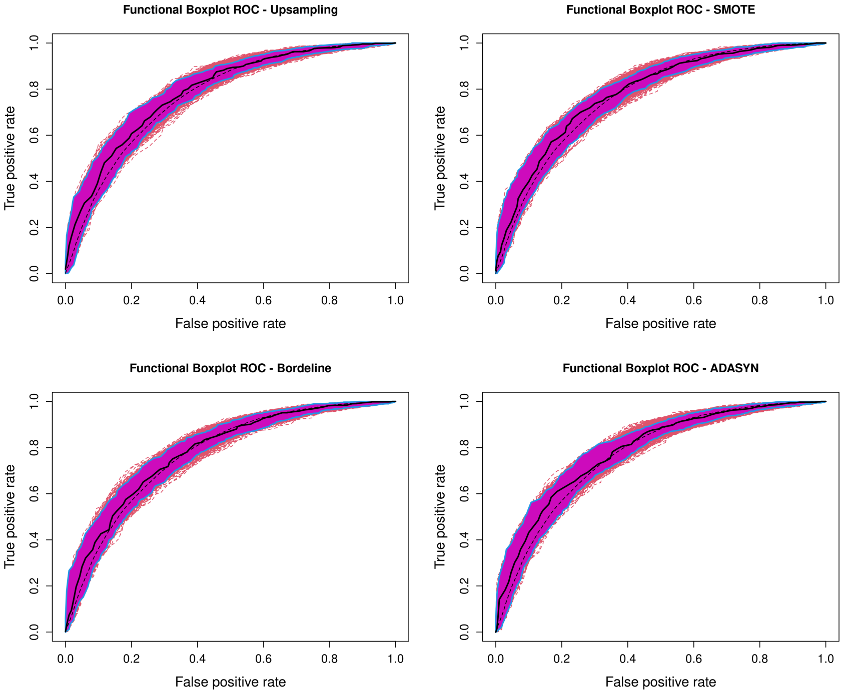

Figure S.11: Functional BoxPlot of the Sentiment Twitter dataset on the train size of 2000

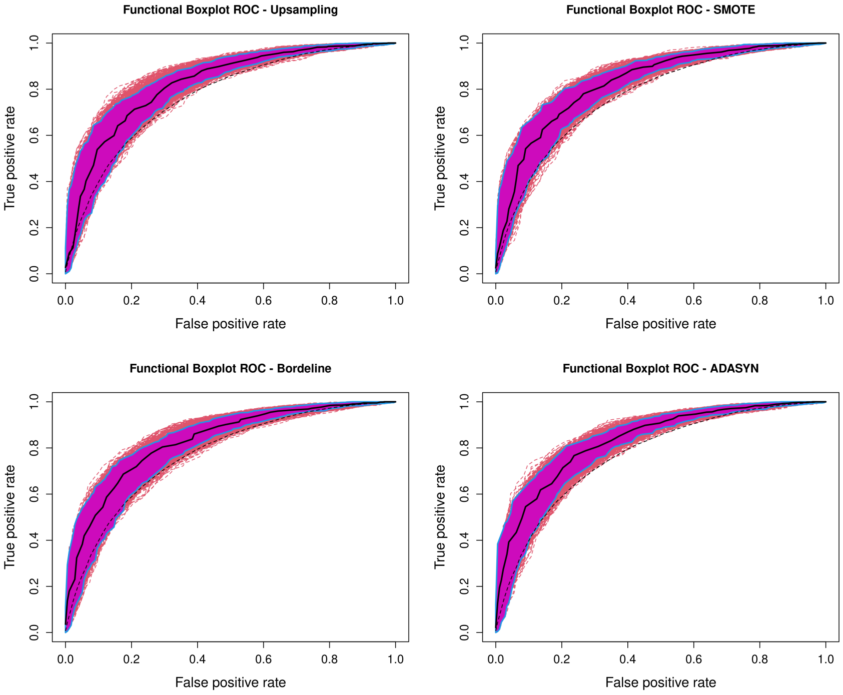

Figure S.12: Functional BoxPlot of the Women's E-Commerce dataset on the train size of 500

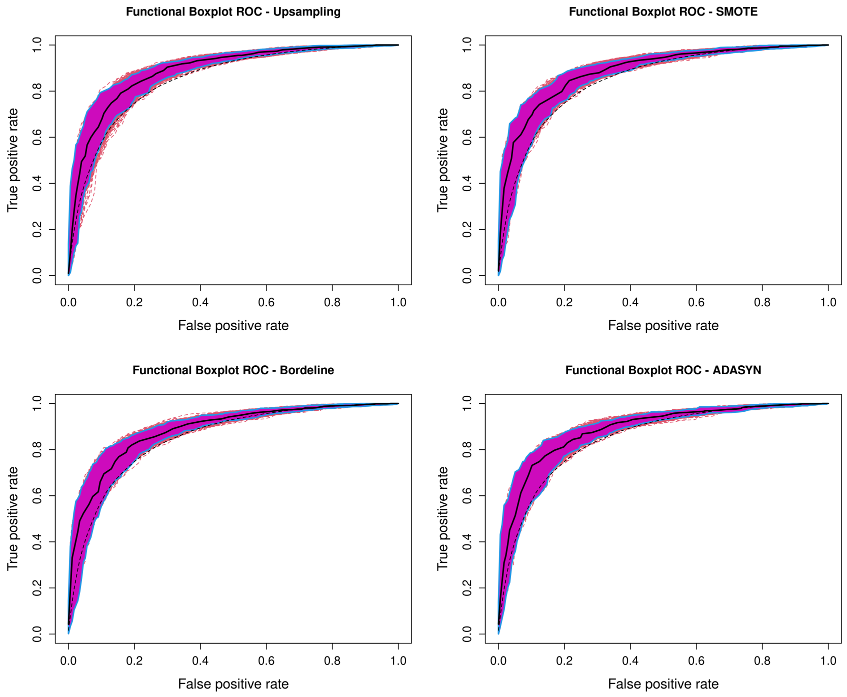

Figure S.13: Functional BoxPlot of the Women's E-Commerce dataset on the train size of 2000

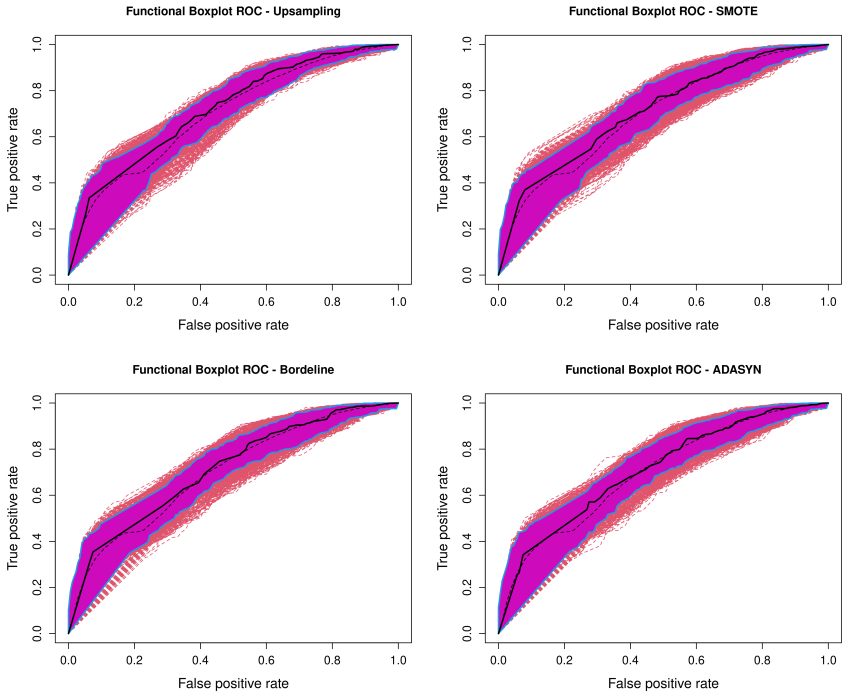

Figure S.14: Functional BoxPlot of the Software review dataset on the train size of 500

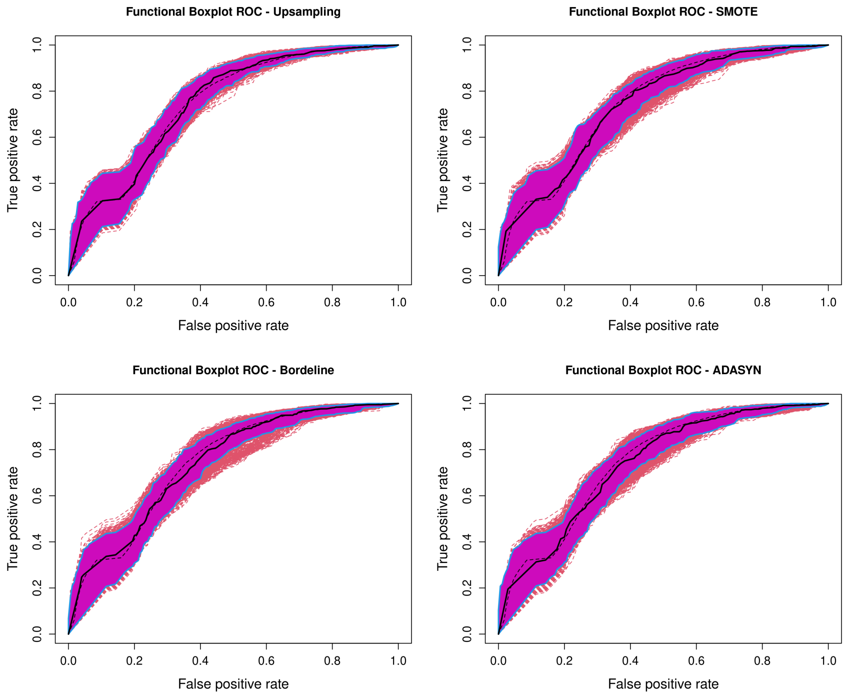

Figure S.15: Functional BoxPlot of the Software review dataset on the train size of 2000

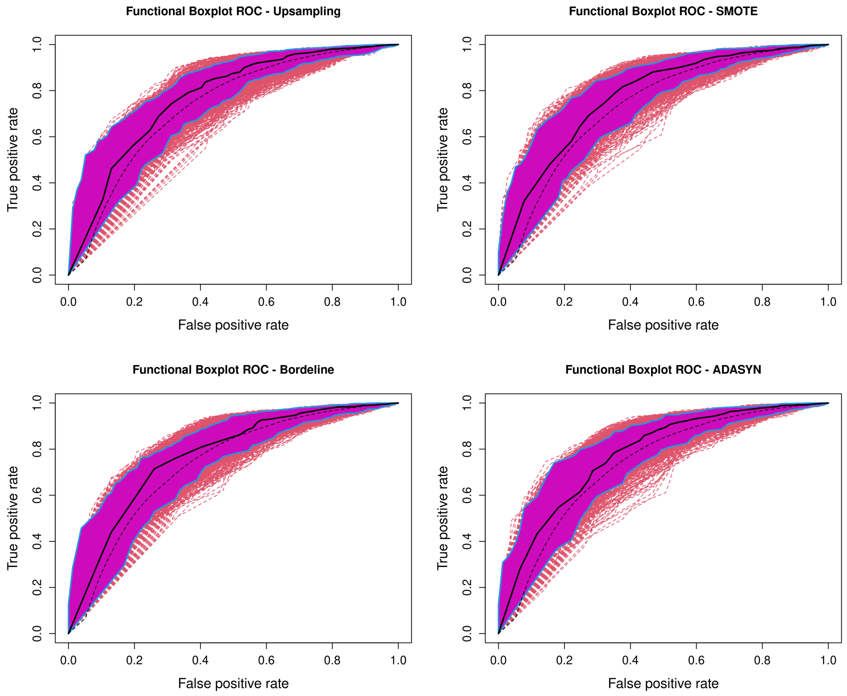

Figure S.16: Functional BoxPlot of the Hate Speech Offensive dataset on the train size of 500

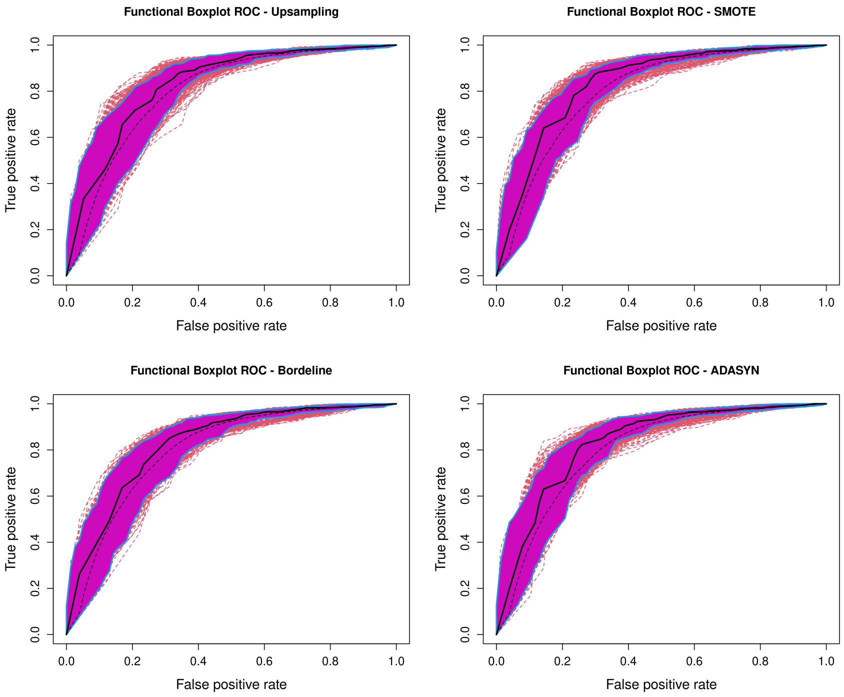

Figure S.17: Functional BoxPlot of the Hate Speech Offensive dataset on the train size of 2000

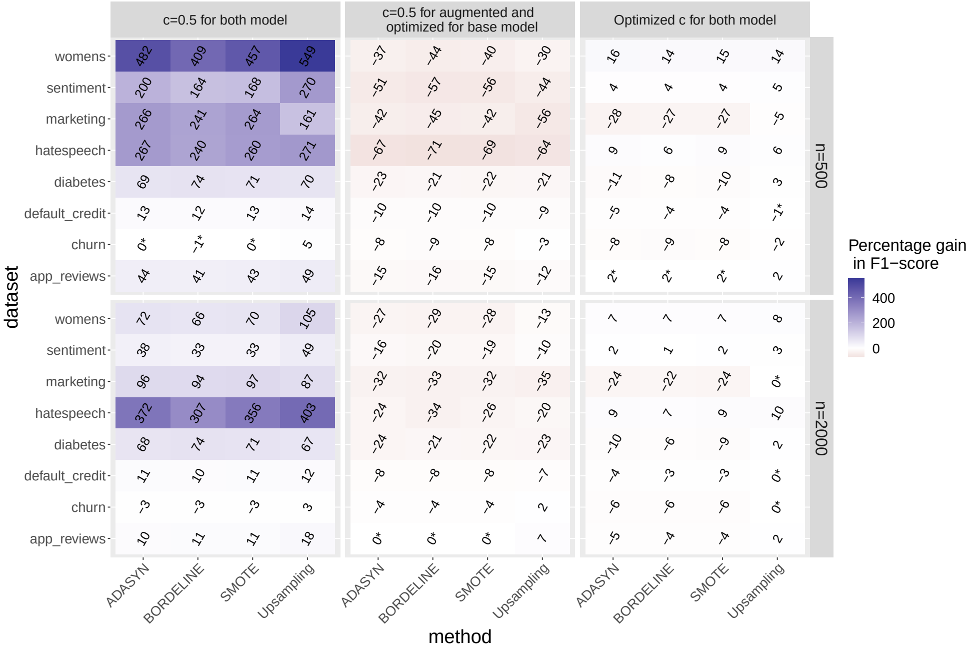

Figure S.18: Heatmap of the mean percentage gain in F1-Score when comparing the augmented methods with the non-augmented model for classification rules. Positive values indicate superior performance by the augmented method. Non-significant gains are marked with asterisks and displayed in white. Our findings indicate a similar result to the balanced accuracy, the data augmentation provides a noticeable benefit only when using the default threshold of $c = 0.5$ (left column); optimizing the threshold on non-augmented data eliminates the need for augmentation.

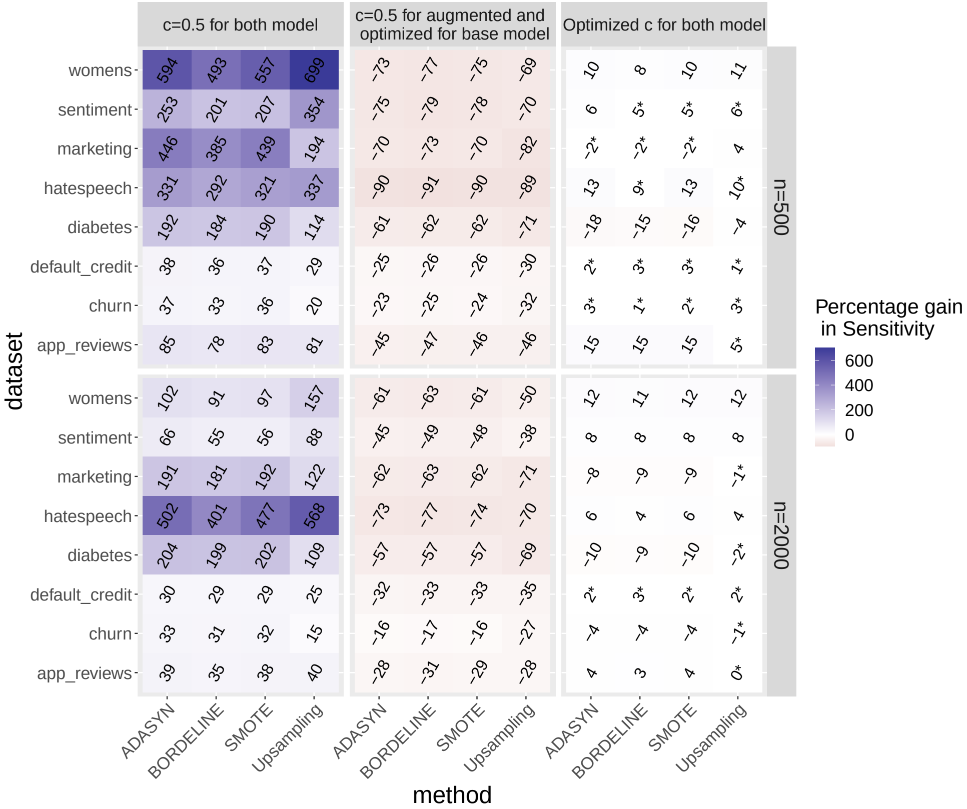

Figure S.19: Heatmap of the mean percentage gain in Sensitivity (minority class) when comparing the augmented methods with the non-augmented model for classification rules. Positive values indicate superior performance by the augmented method. Non-significant gains are marked with asterisks and displayed in white. Our findings for this indicate that optimizing the threshold on non-augmented eliminates the need for augmentation.

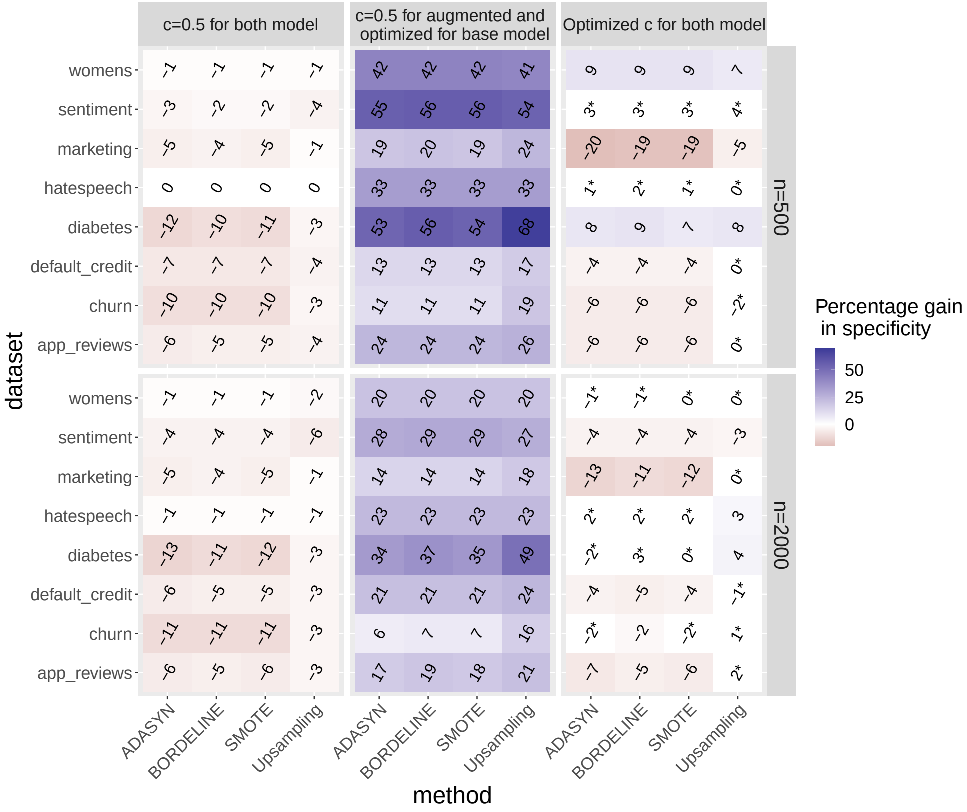

Figure S.20: Heatmap of the mean percentage gain in Specificity (majority class) when comparing the augmented methods with the non-augmented model for classification rules. Positive values indicate superior performance by the augmented method. Non-significant gains are marked with asterisks and displayed in white. Our findings indicate a different behavior for this metric, when using the optimized threshold the augmented method shows an improvement.

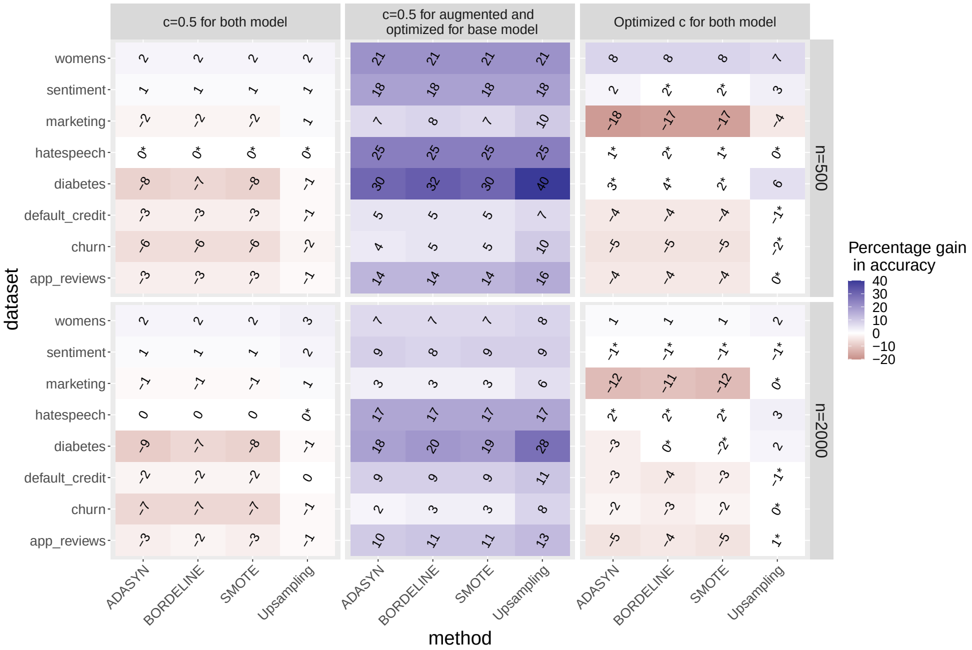

Figure S.21: Heatmap of the mean percentage gain in Accuracy when comparing the augmented methods with the non-augmented model for classification rules. Positive values indicate superior performance by the augmented method. Non-significant gains are marked with asterisks and displayed in white. For this metric when using the default threshold the non-augmented method has a better result, and when optimizing the threshold in a few cases has an increase for the augmented methods.