Using Platt's scaling for calibration after undersampling – limitations and how to address them

Anonymous authors
Paper under double-blind review

Abstract

When modelling data where the response is dichotomous and highly imbalanced, response-based sampling where a subset of the majority class is retained (i.e., undersampling) is often used to create more balanced training datasets prior to modelling. However, the models fit to this undersampled data, which we refer to as base models, generate predictions that are severely biased. There are several calibration methods that can be used to combat this bias, one of which is Platt's scaling. Here, a logistic regression model is used to model the relationship between the base model's original predictions and the response. Despite its popularity for calibrating models after undersampling, Platt's scaling was not designed for this purpose. Our work presents what we believe is the first detailed study focused on the validity of using Platt's scaling to calibrate models after undersampling. We show analytically, as well as via a simulation study, that Platt's scaling should not be used for calibration after undersampling without critical thought. If Platt's scaling would have been able to successfully calibrate the base model had it been trained on the entire dataset (i.e., without undersampling), then Platt's scaling might be appropriate for calibration after undersampling. If this is not the case, we recommend either beta calibration or a modified version of Platt's scaling that fits a logistic generalized additive model to the logit of the base model's predictions, as they are both theoretically motivated and performed relatively well across the settings considered in our study.

1 Introduction

Highly imbalanced, binary classification problems are ubiquitous in today's world of data analytics. These problems appear in fields as varied as finance (e.g., fraud detection; Varmedja et al. 2019), health care (e.g., disease modelling; Shin et al. 2023), and wildfire (e.g., fire occurrence prediction; Phelps & Woolford 2021a). In our study, we will assume that the minority class represents the occurrence of the outcome of interest (denoted by 1; also called the positive class), such as a fraudulent transaction, and the majority class represents its non-occurrence (denoted by 0; also called the negative class). Oftentimes, the datasets associated with imbalanced classification problems are very large (i.e., millions—or more—of observations with many covariates). Both the imbalance and size of these datasets can present modelling challenges. Fitting complex models to massive datasets can be very time-consuming and, in addition, studies have shown that some models tend to neglect the minority class in the face of substantial class imbalance (e.g., Japkowicz & Stephen 2002). A common method for handling both these issues is undersampling, or downsampling (e.g., Wallace & Dahabreh 2014; Moreau et al. 2020; Peng et al. 2020; Phelps & Woolford 2021a; Burmeister et al. 2023; Shin et al. 2023).

When undersampling, all observations from the minority class are kept but only a random subset of the majority class is retained for modelling. The problem with this sampling procedure is that it biases the model. Consider a data distribution $f_{(\mathbf{X},Y)}(\mathbf{x}, y)$, where $\mathbf{x}$ is a vector of covariates and y is a binary outcome, and a model h that produces estimates $h(\mathbf{x}) \approx \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$. We define a perfectly calibrated model as one that generates estimates such that for any $\hat{p} = h(\mathbf{x})$ that can be produced by the model, $\mathbb{P}(Y = 1|h(\mathbf{X}) = \hat{p}) = \hat{p}$, where the probability is computed over the data distribution $f_{(\mathbf{X},Y)}$. For example, for observations where the model assigns $\hat{p} = 0.3$, we expect that 30% are truly positives. Because the distribution of the training data now differs from that of new data, a model h that was well-calibrated on its undersampled dataset will not be well-calibrated when used to make predictions on new data. Generally, a model trained on undersampled data will output estimates that overestimate the true outcome probabilities. This bias induced by undersampling is a serious issue because having poorly calibrated probability estimates can hinder the effectiveness of the model for use in practice, altering prevalence estimates and potentially leading to suboptimal decision-making (e.g., Phelps & Woolford 2021b; Guilbert et al. 2024). Consequently, it is important to adjust the probability estimates of the model to try to obtain well-calibrated estimates whenever possible.

There are several different methods for calibrating models after undersampling. One of the most common approaches is Platt's scaling (Platt 1999). Despite its popularity in this situation (e.g., Wallace & Dahabreh 2014; Moreau et al. 2020; Peng et al. 2020; Phelps & Woolford 2021a; Burmeister et al. 2023; Shin et al. 2023), Platt's scaling was originally designed for another purpose: augmenting the output of support vector machines to obtain calibrated probabilities. It has since been used to calibrate models in other situations, sometimes because models learned from undersampled data (e.g., Wallace & Dahabreh 2014; Moreau et al. 2020; Peng et al. 2020; Phelps & Woolford 2021a; Burmeister et al. 2023; Shin et al. 2023) and sometimes because models were miscalibrated for other reasons (e.g., Guo et al. 2017; Ojeda et al. 2023). Platt's scaling involves fitting a logistic regression model, enforcing a sigmoid relationship between covariates and probabilities. Because of this restriction, the validity of using Platt's scaling outside of its original purpose has been debated; some have criticized it (e.g., Naeini et al. 2015; Kull et al. 2017), but Böken (2021) showed that it is justifiable to use Platt's scaling in more scenarios than other works have suggested. From what we have seen in the literature involving undersampling, it does not appear that the appropriateness of using Platt's scaling in the context of calibration after undersampling is well-understood.

Our work presents what we believe is the first detailed study on the validity of using Platt's scaling to calibrate models after undersampling. First, we analytically show that Platt's scaling is incapable of properly calibrating a model perfectly fit to an undersampled dataset, leading to incorrect estimates of conditional probabilities. We also show how Platt's scaling can be modified so that it can provide correct estimates. Next, we demonstrate that traditional Platt's scaling can be effective when the model fit to undersampled data has a specific systematic error. Finally, we consider another adjustment for improving performance in more general settings.

In Section 2, we outline the calibration approaches considered in this study; Section 3 considers wildland fire occurrence prediction as motivation for our work; Section 4 details the simulation study we conducted; and, in Section 5, we provide conclusions and practical recommendations.

2 Calibration after undersampling

We consider calibrating binary classification models fit to undersampled data. In this context, we have a dichotomous response Y that takes values 0 (i.e., the negative class) or 1 (i.e., the positive class) and associated covariates $\mathbf{X}$, and our goal is to estimate $p(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) \in (0, 1)$. We also introduce a dichotomous variable S that indicates which observations will be used for estimating $p(\mathbf{x})$ by first estimating $\gamma(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}, S = 1) \in (0, 1)$ and then transforming γ using a calibration function κ, giving $p(\mathbf{x}) = \kappa(\gamma(\mathbf{x}))$. Note that if S is independent of $\mathbf{X}$ and Y (e.g., if we take a simple random sample of the training set), then κ is the identity function. However, when undersampling, we keep all positive instances (i.e., $S = 1$ for all positive instances) and negative instances are kept only with probability π_0. We refer to the model that learns from the undersampled training dataset as the base model. If the calibration function κ is learned using another model, we refer to this model as the calibration or secondary model. This approach is an instance of model stacking, whereby a meta learner (the calibration model) is trained based on outputs of one or more base models. The calibration function can also be analytically computed (Dal Pozzolo et al. 2015b). A benefit of analytical calibration is that it does not require learning the calibration function from another full (i.e., not undersampled) dataset, unlike the model-based approaches. However, this is not typically a problem for the model-based approaches because this data is generally available, and it is not computationally expensive to learn this function because it is only based on one covariate. In the next sub-sections, we outline more pros and cons of specific calibration approaches.

2.1 Analytical calibration

A foundational assumption of empirical modelling is that the data used to train the model follows the same distribution as future data (or testing data). Multiple studies have considered the case where this assumption is not met, developing an analytical solution to this problem (e.g., Elkan 2001; Saerens et al. 2002; Dal Pozzolo et al. 2015b). Dal Pozzolo et al. (2015b) considered the specific case where this assumption is not met because of undersampling. Using Bayes' rule, they derived the calibration function shown in Eq. 1.

$$\kappa(\gamma) = \frac{\gamma\pi_0}{1-\gamma+\gamma\pi_0} \tag{1}$$

To calibrate the predictions of a base model fit to an undersampled training dataset, $\hat{\gamma}$, we can use this equation but with the predictions in place of γ. Naturally, if the base model is perfect (i.e., $\hat{\gamma} = \gamma$), then this analytical approach perfectly calibrates the probability estimates. However, this approach does not account for any error in the base model. Therefore, it may struggle in practice, when models are imperfect.

2.2 Platt's scaling and its variations

Platt's scaling (Platt 1999) uses a secondary model to calibrate the predictions of the base model. Platt's scaling involves fitting a logistic regression model to the y's using the $\hat{\gamma}$'s as the predictor. In the original paper, Platt (1999) slightly modified the responses to perform regularization, but Platt's scaling has often been implemented without this (e.g., Phelps & Woolford 2021a; Ojeda et al. 2023). It is therefore sometimes called logistic calibration (e.g., Kull et al. 2017), although some still make a distinction between the two (Ojeda et al. 2023). In our work, we leave the responses as 0/1 variables. When using logistic regression, we assume that $Y \sim \text{Bernoulli}(p)$ and that there is a linear relationship between the logit of the p's and the $\hat{\gamma}$'s (see Eq. 2).

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\hat{\gamma} \tag{2}$$

After fitting the logistic regression model (i.e., learning estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for β_0 and β_1, respectively), we obtain the following calibration model from this approach:

$$\kappa(\hat{\gamma}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1\hat{\gamma})}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1\hat{\gamma})} \tag{3}$$

To assess the validity of Platt's scaling as a method for calibration after undersampling, we must determine if the relationship assumed by this approach is reasonable. Naturally, this depends on the base model, since it outputs the $\hat{\gamma}$'s. We first consider the case where the base model is perfect, resulting in Theorem 1.

Theorem 1. *If the base model provides perfect estimates based on the undersampled training dataset, Platt's scaling is incorrectly specified and cannot provide perfect probability estimates on the full dataset.*

Proof: Since we are considering a perfect base model, we can substitute Eq. 1 for p into Eq. 2 to determine the true relationship between p and $\gamma = \hat{\gamma}$ on the log odds scale. After some algebra, the result of this substitution is Eq. 4.

$$\log\left(\frac{p}{1-p}\right) = \log(\pi_0) + \log\left(\frac{\gamma}{1-\gamma}\right) \tag{4}$$

Clearly, the relationship between the p's and the γ's is not linear on the log odds scale. Thus, even though the base model perfectly modeled its training dataset, the secondary model that will be learned from Platt's scaling cannot possibly properly adjust the $\hat{\gamma}$'s to achieve a perfect final model. ☐

For those familiar with both the literature on Platt's scaling and on undersampling, the result of Theorem 1 might be expected. One of the criticisms of Platt's scaling is its inability to leave a perfect model perfectly calibrated (e.g., Kull et al. 2017), and logistic models trained on undersampled data are calibrated through only adjusting their intercept (e.g., Taylor et al. 2013; Phelps & Woolford 2021a), so it is intuitive that Platt's scaling is unable to calibrate a model that is perfect with respect to the undersampled data. However, Platt's scaling's inability to calibrate such models is more disguised in this setting, as Platt's scaling generally will still improve calibration due to the base model's extreme overprediction because of learning from the undersampled dataset (e.g., Phelps & Woolford 2021b).

Theorem 1 shows that Platt's scaling cannot properly calibrate a perfect base model, but a corollary of Eq. 4 is that a simple transformation can be done to remedy this problem. Rather than using $\hat{\gamma}$ as the covariate, we can use $\log\{\hat{\gamma}/(1-\hat{\gamma})\}$. This transformation has been considered in the calibration literature before, such as by Turner et al. (2014) for linear in log odds calibration, by Leathart et al. (2017) for fitting probability calibration trees (which involve fitting a logistic regression model in the leaf nodes), and by Kull et al. (2017), who call this beta$[a = b]$ calibration (see the next sub-section for more details). Böken (2021) pointed out that Platt's scaling was designed for a predictor that can take any real value, so it makes sense to use this transformation to convert predictions from $[0, 1]$ to the real line. Via their simulation study (which did not incorporate undersampling), Ojeda et al. (2023) found that using the logit transformation generally improved calibration. Although this approach has appeared several times in the calibration literature, we are not aware of any work that has explicitly theoretically motivated its use for calibration after undersampling. Of course, we cannot expect the base model to be perfect in practice, but we expect that defining the logistic regression model using this transformation of $\hat{\gamma}$ will still be effective for models where $\hat{\gamma} \approx \gamma$.

Models sometimes exhibit systematic estimation errors (e.g., Niculescu-Mizil & Caruana 2005; Guo et al. 2017; Guilbert et al. 2024). Some models, such as random forests and boosted decision trees, have been shown to push probability estimates towards 0.5, while models like Naïve Bayes push estimates towards extreme values (i.e., 0 or 1). We now consider models where $\hat{\gamma}$ and γ have a sigmoidal relationship, represented by Eq. 5.

$$\gamma = \frac{1}{1 + \exp[-r(\hat{\gamma} - m)]} \tag{5}$$

Here, r and m are arbitrary constants. Mathematically, we require only that $r \in \mathbb{R}$ and $m \in [0, 1]$, but it is worth noting that settings within these bounds may not lead to a reasonable representation of a model. Under settings that reasonably represent a model (e.g., $r = 10$ and $m = 0.5$), this relationship represents a model whose estimates are pushed towards 0.5. Calibrating models that err in the form indicated by Eq. 5 is considered a valid use of Platt's scaling (e.g., Kull et al. 2017; Leathart et al. 2017) because the logistic regression model's parametric assumptions are met. In Theorem 2, we show that these assumptions are still met when the base model learns from an undersampled dataset.

Theorem 2. *If the base model provides predictions that have a sigmoidal relationship with the true probabilities on the undersampled training dataset, Platt's scaling is correctly specified and can provide perfect probability estimates on the full dataset.*

Proof: We can substitute Eq. 5 for γ in Eq. 4. After some algebra, we can obtain Eq. 6.

$$\log\left(\frac{p}{1-p}\right) = \log(\pi_0) - rm + r\hat{\gamma} \tag{6}$$

The relationship between the logit of the p's and the $\hat{\gamma}$'s is now linear, so the assumption of the logistic regression model is met. A logistic regression model with the learned coefficients $\hat{\beta}_0 = \log(\pi_0) - rm$ and $\hat{\beta}_1 = r$ would perfectly calibrate this base model. □

We have addressed how models can be calibrated after undersampling if the model is perfect or if its predictions have a sigmoid shape with the true probabilities, but we have not addressed models with a tendency to push estimates towards extreme values or models that deviate somewhat from either perfect prediction or a perfect sigmoid shape. In both cases, we cannot derive a simple transformation that will allow us to satisfy the assumptions of logistic regression. However, this does not preclude us from being able to modify Platt's scaling to obtain better probability estimates. Rather than instituting the restrictive assumptions of logistic regression, a logistic generalized additive model (GAM) can be used. This has also been used for calibration before (e.g., Lucena 2018), including to calibrate models after undersampling in a few studies (e.g., Coussement & Buckinx 2011; Phelps & Woolford 2021a), but it does not seem to be a common approach for this purpose. Logistic GAMs use smoothers to model the relationship between covariates and the outcome, relaxing the assumption of linearity on the log odds scale so that non-linear relationships can be modeled. It is important to note that, given enough data, a logistic GAM will converge to a logistic regression model when the linearity assumption holds.

2.3 Beta calibration

As mentioned previously, the primary criticism of Platt's scaling is its restrictive parametric assumptions. Consequently, Kull et al. (2017) developed beta calibration. Beta calibration is still a parametric method, but it is more flexible than Platt's scaling. The calibration function learned in beta calibration is shown in Eq. 7. We have adopted the parameter names (i.e., a, b, and c) of Kull et al. (2017) but have expressed the calibration function differently to make it more comparable to Platt's scaling as presented herein.

$$\kappa(\hat{\gamma}) = \frac{\exp[c + a\log(\hat{\gamma}) - b\log(1-\hat{\gamma})]}{1 + \exp[c + a\log(\hat{\gamma}) - b\log(1-\hat{\gamma})]} \tag{7}$$

A special case of beta calibration is beta[$a = b$] calibration. As mentioned in Section 2.3, this is equivalent to Platt's scaling with the logit transformation with $\hat{\beta}_0 = c$ and $\hat{\beta}_1 = a = b$. Throughout our study, we will refer to Platt's scaling with the logit transformation as a variation of Platt's scaling, but it should be noted that it could also be referred to as a variation of beta calibration.

Kull et al. (2017) showed that beta calibration performed similarly to Platt's scaling when Platt's scaling's assumptions were approximately met (i.e., when calibrating models that pushed probability estimates towards 0.5). When the assumptions of Platt's scaling were not met, beta calibration clearly outperformed Platt's scaling. However, that study did not focus on calibrating models after undersampling.

2.4 Isotonic regression

Isotonic regression (e.g., Zadrozny & Elkan 2002) is the most flexible calibration method we consider in this study. It minimizes $\sum_i [y_i - \kappa(\hat{\gamma}_i)]^2$, where κ is a step function. This model is learned using the pair-adjacent violators algorithm (Ayer et al. 1955), and the only restriction imposed on κ is that it must be monotonically non-decreasing. This is a very flexible method, but because κ is a step function it does not produce a smooth relationship between the base model's predictions and the new probability estimates.

3 A motivating example: Wildland fire occurrence prediction

Wildland fire occurrence prediction is an important part of a fire management agency's planning. Often, this prediction is done by mapping space-time into voxels such that it is reasonable to model fire occurrences as a presence/absence problem. This results in an extremely imbalanced binary classification problem. For example, de Haan-Ward et al. (2024) had less than 0.1% positive cases in their study region of the province of Ontario. We also consider data from Ontario, provided to us by the Ontario Ministry of Natural Resources and spanning the years 2000 to 2019. Using 20km × 20km × daily voxels, the dataset provides information on the day of year and location, the weather, including both standard weather variables (e.g., temperature) and fire-weather variables (e.g., Fine Fuel Moisture Code), as well as variables describing land use. To avoid the effects of a potential temporal trend, we split the data as follows: years 2000, 2003, 2006, 2009, 2010, 2013, 2016, and 2019 were used for training; years 2001, 2004, 2007, 2011, 2014, and 2017 were used for calibration; and years 2002, 2005, 2008, 2012, 2015, and 2018 were used for testing. We would not recommend this splitting for operational use, where detecting temporal trends would be important.

For our example, our base model is a logistic GAM that predicts the occurrence of human-caused fires using smoothers for day of year, temperature, relative humidity, wind speed, rain, Fine Fuel Moisture Code, area of infrastructure interface, urban interface, and industrial interface, as well as longitude and latitude (as a bivariate smoother). To create the training dataset, we kept all fire observations and 0.2% of the non-fire observations, leading to just over a quarter of the observations being fires. A GAM is useful for illustrative purposes in this setting because we can compare the predictions obtained from various calibration approaches to the predictions obtained from the well-established method of adjusting the intercept of the GAM (e.g., Taylor et al. 2013; Phelps & Woolford 2021a). Adding an offset of $\log(\pi_0)$ to the intercept learned from the undersampled dataset has been shown to account for the bias induced by the sampling procedure. Note that this offset is equivalent to using analytical calibration for the GAM.

The calibration methods we considered here are Platt's scaling and Platt's scaling with the logit transformation. Reliability plots for both of these approaches are shown in Fig. 1. These were constructed by creating bins of width 0.001 based on the predictions, then computing the average prediction and average rate of fire occurrence within each bin and plotting them. We plotted only the bins with more than 50 observations to avoid extremely noisy results. We also added 95% prediction intervals, computed assuming the predicted probabilities were correct. The results seem to show that Platt's scaling underestimates the probabilities for relatively large probabilities. When using the logit transformation, this systematic underestimation appears to go away, although it is difficult to make definitive conclusions because of the noisiness in the data. It should be noted that plotting only bins with more than 50 observations led to removing the 48 observations with an estimate greater than 0.023 for traditional Platt's scaling (none of which were truly fires) and removing the 362 observations with an estimate greater than 0.037 for Platt's scaling with the logit transformation. The mean prediction for these observations was 0.047, while the true mean occurrence was 0.044.

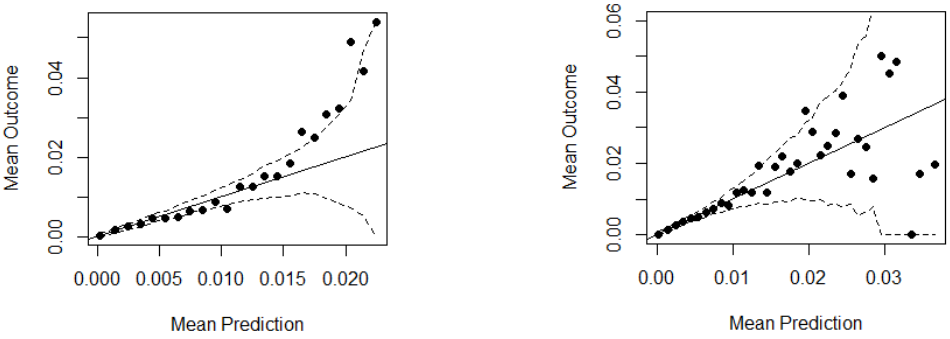

Figure 1: Reliability plots for a logistic generalized additive model (GAM) calibrated using Platt's scaling (left) and Platt's scaling with the logit transformation (right). The solid line is the 45° line and the dashed lines represent 95% prediction intervals, computed assuming the predicted probabilities from the modelling were correct.

Fig. 2 has line plots comparing the predictions of the two Platt's scaling approaches to the predictions from the model with an offset. The plots show that Platt's scaling and the offset lead to very different predictions; the offset generates much larger predictions. This is consistent with the underprediction observed in the reliability plot. On the other hand, the predictions from Platt's scaling with the logit transformation and the offset are nearly identical, suggesting that the logit transformation has worked well.

Although it appears that Platt's scaling leads to underestimates of the probability of a fire for the higher probability cases and that this may be fixed by using the logit transformation, it is difficult to make definitive conclusions. The rightmost points on the reliability plots are subject to high variability, so our conclusion that the logit transformation has helped is based largely on how well the predictions from that approach align with the predictions from using the offset. Thus, a simulation study can provide useful insights here. We can simulate as much data as needed and, more importantly, the true probabilities would be known, so we need not rely on reliability plots or comparisons with other methods—we can directly compare the estimates from each method to the true probabilities.

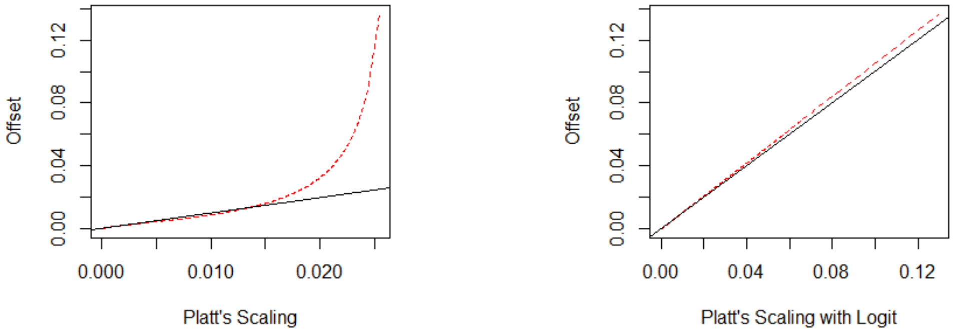

Figure 2: Plots comparing the predictions of the model with an offset to the model without an offset calibrated using Platt's scaling (left) and Platt's scaling with the logit transformation (right). The solid line is the 45° line and the dashed line shows the relationship between the predictions.

4 Simulation study

4.1 Methods

In this section, we describe the simulation study we used to evaluate the four variations of Platt's scaling outlined in Section 2.2: traditional Platt's scaling, Platt's scaling with the logit transformation, Platt's scaling with a logistic GAM, and Platt's scaling with both the logit transformation and a logistic GAM. These methods were compared to analytical calibration, beta calibration, and isotonic regression. Our simulation study was implemented in R (R Core Team 2023).

4.1.1 Simulating data

To be able to compare different calibration methods, we simulated datasets with known outcome probabilities. We defined a success as belonging to the positive class and consider three different data generating processes, each with different mean probabilities of success. This was done so that we could study the effects of different levels of class imbalance. Each data generating process creates data that are imbalanced in terms of the response variable, making them amenable to undersampling. The mean success probabilities are approximately 0.0022, 0.0208, and 0.1109. Specific details about how data were generated are given in Appendix 1. We created calibration datasets with 100 000 and 1 000 000 observations so that we could investigate the effectiveness of the methods with varying amounts of data. Our testing dataset had 1 000 000 observations.

4.1.2 Simulating model outputs

Rather than modelling the output of our simulated datasets as a function of the covariates, we defined hypothetical base models with various estimation errors. The first hypothetical base model we considered is a perfect model (i.e., $\hat{\gamma} = \gamma$). Note that this model still required calibration to account for the undersampling process. Next, we considered models whose generated $\hat{\gamma}$'s systematically deviate from the γ's. As discussed previously, systematic error has been found in very well-known models; random forests and boosted trees have been shown to push probability estimates towards 0.5, while Naïve Bayes models have been shown to push estimates towards extreme values (i.e., 0 or 1) (e.g., Niculescu-Mizil & Caruana 2005; Guilbert et al. 2024). We have represented models that push probability estimates towards 0.5 with the relationship shown in Eq. 8. Models that push probability estimates towards extreme values are represented by the relationship in Eq. 9. Plots illustrating each of these relationships are available in Appendix 1.

$$\hat{\gamma} = \min\left[\max\left(-\frac{1}{10}\log\left(\frac{1}{\gamma} - 1\right) + 0.5,\, 0\right),\, 1\right] \tag{8}$$

$$\hat{\gamma} = \frac{1}{1 + \exp[-10(\gamma - 0.5)]} \tag{9}$$

Finally, we considered a model that generates nearly perfect predictions in expectation, but does at times make larger errors. This was implemented by incorporating noise on the log odds scale of the base model. This noise was added via a normally distributed random variable with a mean of zero and standard deviation of 0.2. Because the datasets are imbalanced, this results in $\hat{\gamma}$'s with a mean slightly larger than the mean of the γ's. A scatterplot showing the relationship between the $\hat{\gamma}$'s and the γ's when the mean outcome probability is 0.0208 is shown in Appendix 1.

To obtain the γ's for each data generating process, we set sampling rates that would generate approximately balanced training datasets. For the data generating processes described in Section 4.1.1, we used sampling rates of $\pi_0 = 0.0023$, $\pi_0 = 0.02125$, and $\pi_0 = 0.125$.

4.1.3 Implementing and evaluating the calibration methods

Analytical calibration was implemented using only base R. To implement the four variations of Platt's scaling, we used the glm and gam functions from the stats and mgcv packages (Wood 2011), respectively. Beta calibration was implemented using the betacal package (Kull et al. 2021). Like Platt's scaling, isotonic regression was implemented using functions from the stats package. For each isotonic regression model, we fit the model then converted it to a step function to make predictions on the testing dataset.

In our study, we paid special attention to the ability of each calibration method to fit the true relationship between the $\hat{\gamma}$'s and the p's. Unlike when working with real data, the p's are known in our study. We can take advantage of this by creating line plots of the $\hat{p}$'s against the p's to visually evaluate the calibration of the predictions across the entire spectrum of probabilities. For the model whose predictions are nearly perfect in expectation, line plots do not lead to a clear curve because of the randomness in the model's errors. Instead, we created reliability plots, but used the true probabilities instead of the outcomes (as is typically done in practice with real data) to eliminate the effects of noise in the outcome. Since we used the probabilities, we plotted bins with more than 10 observations (as opposed to more than 50 in Section 2). To avoid relying entirely on visual assessments, we also measured the gap between the $\hat{p}$'s and the p's using root mean squared error (RMSE) and mean absolute error (MAE).

4.2 Results

4.2.1 Perfect base model

The line plots for each of our simulation settings revealed similar results, so we only show two representative plots (Fig. 3). All six figures are available in Appendix 2. Here, we omit analytical calibration because it is known that it will perfectly calibrate the predictions.

In Fig. 3, the lack of smoothness in the isotonic regression is very apparent. Although the model is flexible, its performance was quite poor, especially for the rarer cases with relatively high probabilities. Likewise, Platt's scaling struggled with these cases, systematically producing estimates that were far too low. These results confirm our analysis in Section 2.2, showing that Platt's scaling is unable to properly calibrate a perfect base model after undersampling. They also provide additional information about the way in which Platt's scaling fails in this setting: severe underestimation of success for higher probability outcomes. Using the logit transformation, however, fixed this problem. With sufficient data, both Platt's scaling approaches with this transformation yielded nearly perfect predictions. With less information (i.e., through a smaller calibration dataset and smaller success rate), using the GAM offered worse performance because of its additional flexibility (see Appendix 2). Beta calibration also performed similarly to these two models. Although there is some deviation in the predictions from these three models with a calibration dataset of 100 000 observations, their predictions were nearly identical with a calibration dataset of 1 000 000 observations.

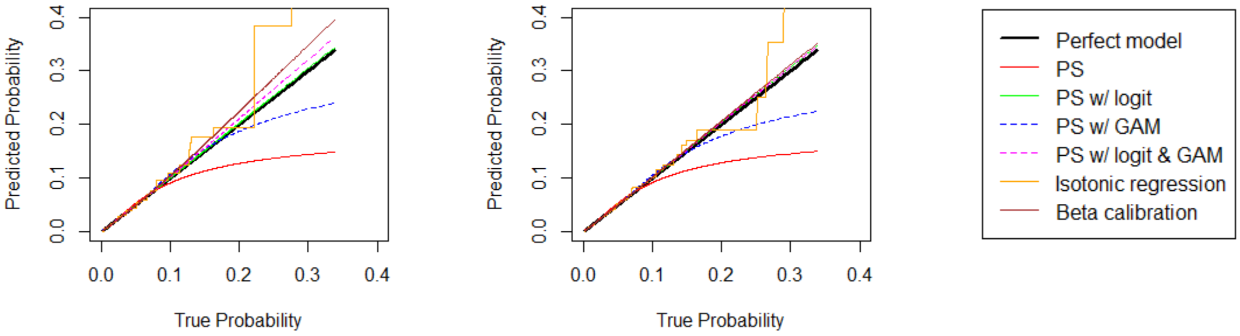

Figure 3: Probability estimates on the testing dataset from Platt's scaling (PS) and its variations, as well as isotonic regression and beta calibration. The data generating process has a mean probability of success of 0.0208. The base model is perfect and the calibration methods were trained on 100 000 observations (left) and 1 000 000 observations (right).

4.2.2 Base model pushes probability estimates towards 0.5

When the base model pushes probability estimates towards 0.5, the relative effectiveness of the calibration methods changed considerably (see Fig. 4 and additional results in Appendix 3). Notably, analytical calibration performed incredibly poorly because it does not account for the miscalibration of the base model at all. However, traditional Platt's scaling does account for this miscalibration, as shown in Section 2.2, resulting in it generally being the most effective approach with this base model. Like with a perfect base model, Platt's scaling with a GAM was just as effective if the mean outcome success rate or size of the calibration dataset were sufficiently large. Once again, beta calibration performed similarly to the two Platt's scaling approaches with the logit transformation, with all of them generally overestimating the probability of success for higher probability outcomes. Of those two Platt's scaling approaches, using the GAM provided improved performance in the datasets with 1 000 000 observations (see Appendix 3). One unexpected result was that Platt's scaling with the logit transformation (with or without the GAM) performed best in terms of RMSE when the mean outcome success rate was 0.0022 and 100 000 observations were in the calibration dataset. This may simply be due to randomness in the simulation process, however, as traditional Platt's scaling outperformed Platt's scaling with the logit transformation in all other cases, including when the mean outcome success rate was the same but more observations were used for calibration.

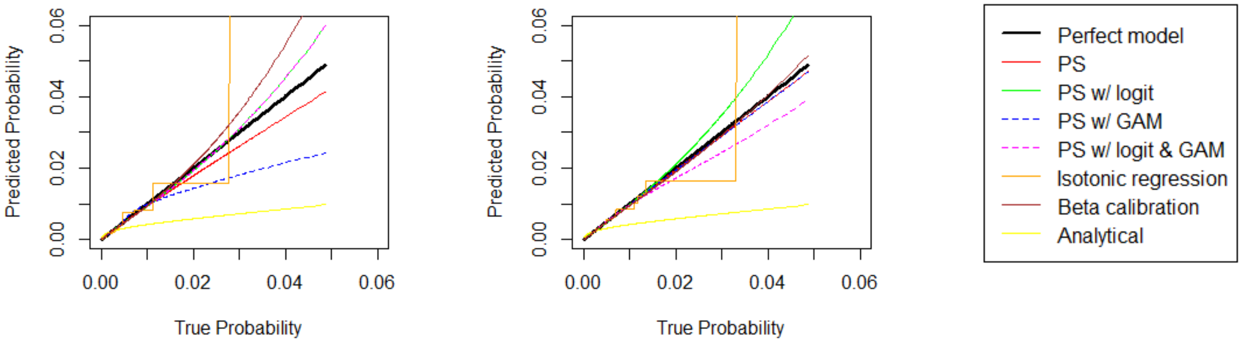

Figure 4: Probability estimates on the testing dataset from Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration. The data generating process has a mean probability of success of 0.0022. The base model pushes probability estimates towards 0.5 and the calibration methods were trained on 100 000 observations (left) and 1 000 000 observations (right).

4.2.3 Base model pushes probability estimates towards 0 or 1

When the base model pushes probability estimates towards extreme values, none of the calibration methods generated predictions that closely aligned with the true probabilities (see Fig. 5 and additional results in Appendix 4). Except for analytical calibration and isotonic regression, all the methods underestimated success probabilities for higher probability observations. Analytical calibration severely overestimated these probabilities and provided terrible estimates. Isotonic regression is the only calibration method that did not exhibit visible systematic biases with this base model. However, its use of a step function still limits how well it can approximate the true probabilities. In terms of MAE and RMSE (see Appendix 4), Platt's scaling using the logit transformation and a GAM generally performed the best. This method was only beaten twice, once by both beta calibration and Platt's scaling with the logit transformation and once by only beta calibration. In both cases, the mean outcome success rate was 0.0022. Although Platt's scaling using the logit transformation and a GAM did systematically underestimate success probaiblities for higher probability events, its underestimation was usually less severe than the other Platt's scaling approaches and beta calibration. The exception to this statement is when the mean outcome success rate was 0.0022, where beta calibration's underestimation was less severe. For larger success rates and calibration datasets, isotonic regression was the second-best performing calibration method.

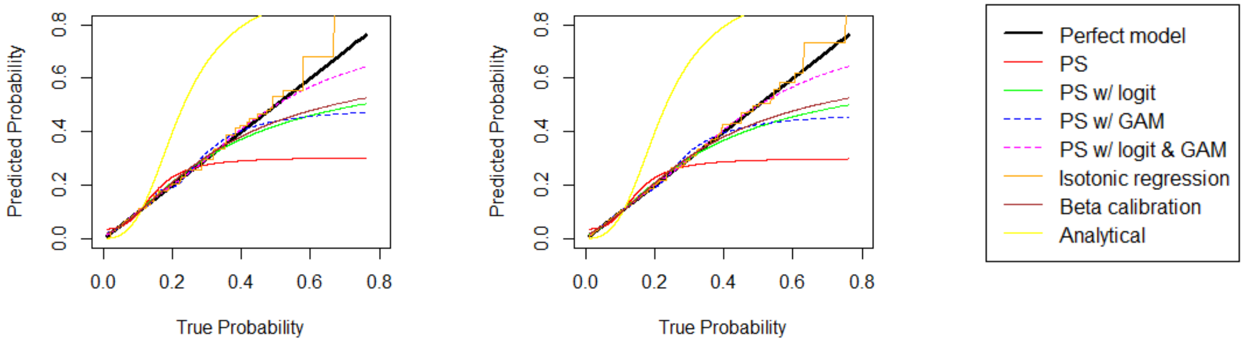

Figure 5: Probability estimates on the testing dataset from Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration. The data generating process has a mean probability of success of 0.1109. The base model pushes probability estimates towards 0 or 1 and the calibration methods were trained on 100 000 observations (left) and 1 000 000 observations (right).

4.2.4 Base model is nearly perfect in expectation

For the base model that is nearly perfect in expectation, the line plots we've used thus far do not work because of the noise in the model's errors. Consequently, we created reliability plots instead. Note that these plots must be interpreted differently from the plots in Figs. 3 - 5. If the line is directly along the 45° line, it no longer means the calibration method generates correct probability estimates. Instead, it means that the model is perfectly calibrated. Recall that this means that $\mathbb{P}(Y = 1|h(\mathbf{X}) = \hat{p}) = \hat{p}$ for all $\hat{p}$, where h is the model. The axes for reliability plots also differ from the axes in the earlier figures. With sufficient information, it seems that nearly all of the calibration methods lead to fairly well-calibrated predictions. The exception to this is Platt's scaling, which consistently produced poorly calibrated predictions. Using a GAM helped except for when the mean outcome success rate was 0.0022 (see Appendix 5). It should be noted that isotonic regression had some sets of larger predictions that are not shown in the plots.

Since the plots in Fig. 6 provide less information, we relied more on the quantitative assessment to determine how well each calibration method worked in this setting (see Appendix 5). When the base model is nearly perfect in expectation, Platt's scaling with the logit transformation was generally the most effective calibration method. Analytical calibration was no longer perfect because of the errors in the base model's predictions, but it also did not perform nearly as poorly as it did when the base model pushed probability estimates towards 0.5 or towards extreme values. When the mean outcome success rate was 0.0022 and

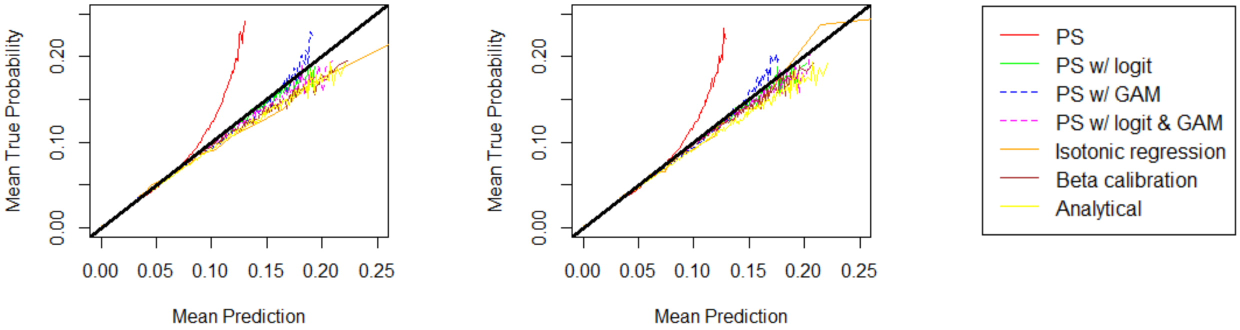

Figure 6: Reliability plots for Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration. The data generating process has a mean probability of success of 0.0208. The base model is nearly perfect in expectation and the calibration methods were trained on 100 000 observations (left) and 1 000 000 observations (right).

the calibration dataset had 100 000 observations, analytical calibration was competitive with Platt's scaling with the logit transformation as the best method. However, as the other calibration methods were given more information, either through increasing the size of the calibration dataset or through the mean outcome success rate increasing, the relative effectiveness of analytical calibration diminished. These results again reinforce why Platt's scaling should not be blindly used to calibrate models after undersampling; it often performed considerably worse than Platt's scaling with the logit transformation.

4.3 Discussion

Our simulation study has shown that the effectiveness of different calibration methods varies substantially based on a number of factors. Although we have considered the level of class imbalance in the data and the size of the calibration dataset, our primary consideration was the performance of the base model. Our results show that this is an extremely important factor to consider. While our simulation study has shown that traditional Platt's scaling can perform poorly in some cases, we also found that one of the four variations of Platt's scaling was the best-performing calibration method in all cases except with the perfect base model. Thus, Platt's scaling or a variation of it can be a valuable tool for calibrating models after undersampling.

Although traditional Platt's scaling performed poorly when the base model was perfect, the variation of Platt's scaling with the logit transformation performed well. It did not perform as well as analytical calibration because that method perfectly adjusts for undersampling in this setting. However, analytical calibration is risky to use because it can perform terribly with imperfect models. In addition, even with a model that is nearly perfect in expectation, Platt's scaling with the logit transformation was able to outperform it.

Traditional Platt's scaling was not effective when the base model was perfect, but it worked very well when the base model pushed probabilities towards 0.5. In this setting, the relationship between the model's outputs and the true probabilities is linear on the log odds scale, so the assumptions of the logistic regression model are met. Thus, even though Platt's scaling is unable to properly calibrate a perfect base model after undersampling, it might be suitable sometimes.

When the base model pushes probabilities towards 0 or 1, none of the calibration methods worked very well, but Platt's scaling with both the logit transformation and a GAM tended to lead to the best-calibrated probabilities. This is one of the safest calibration choices, as it was generally in the top three based on RMSE and MAE regardless of the base model. Using Platt's scaling with a GAM, whether with the logit transformation or not, seems to be a reasonable choice because of its relatively strong performance and robustness to the base model. However, this approach was sometimes still considerably less effective than one of the other Platt's scaling approaches when the outcome probability and calibration dataset size were both relatively small. Similar comments can be made about isotonic regression, whose use led to identical probability estimates regardless of the base model used (excluding the results in Section 4.2.4, which involve randomness). However, although isotonic regression was robust to the base model (even more so than Platt's scaling with a GAM), its performance generally was not very good.

Our simulation study has also shown that the use of Platt's scaling can have undesirable effects on model selection. Consider the case where we fit several base models and would like to compare them so that we can choose a final model. In our simulation study, if we use traditional Platt's scaling to calibrate each model, then we would choose the model that pushes probabilities towards 0.5 as our final model, even though one of our candidate models is perfect. This could negatively affect our ability to interpret the model as we try to better understand the underlying data generating process.

5 Conclusion

In this paper, we have shown that Platt's scaling is generally not a good choice for calibrating models trained on an undersampled dataset. Although it can work, Platt's scaling relies on the base model having a specific systematic error to properly calibrate the predictions. With a perfect base model, calibration via Platt's scaling results in predictions that underestimate success probabilities for higher probability outcomes. For a field such as wildland fire management, this could cause substantial problems due to underestimation of wildland fires. However, a modified version of Platt's scaling based on the logit of the base model's predictions is an effective calibration approach with a perfect base model. This is also known as beta$[a = b]$ calibration. Using beta calibration (without the restriction that $a = b$) or a logistic GAM instead of a logistic regression model can also lead to improved calibration.

To choose a calibration method in practice, the most robust approach is to compare the different methods and then choose the best one. A practitioner could qualitatively and quantitatively evaluate each method given a particular base model, and then select the best calibration method. For real data, the true probabilities are unknown, so metrics like Brier score or negative logarithmic score would be needed in place of RMSE and MAE. Of the two, we recommend NLS because it can better reflect differences in the models (Benedetti 2010). Other metrics may also be viable options, like customized metrics from the Beta family of scoring rules (Merkle & Steyvers 2013). It is also important to note that to obtain an unbiased estimate of the performance of the entire modelling procedure, an additional dataset is needed. Using the metrics obtained on the testing dataset (which were used to choose the calibration method) will result in a biased estimate.

This quantitative comparison is a relatively time-consuming process, so a practitioner may wish to bypass this procedure. This may be possible by critically thinking about the base model being used. In general, our results indicate that a practitioner should only use traditional Platt's scaling for calibration after undersampling if it would have been able to calibrate the base model had it been trained on the entire dataset (i.e., without undersampling) (e.g., see Böken 2021). For example, boosted trees and random forests tend to push probability estimates towards 0.5 (e.g., Niculescu-Mizil & Caruana 2005; Guilbert et al. 2024), so one might choose to use Platt's scaling for calibration anyways when using these models. Traditional Platt's scaling can simultaneously account for a base model pushing its estimates towards 0.5 and being miscalibrated due to undersampling, so it might be a good choice in this situation. However, this justification for using Platt's scaling is generally not given in the undersampling literature; oftentimes, the only reasoning given is miscalibration due to undersampling (e.g., Wallace & Dahabreh 2014; Moreau et al. 2020; Peng et al. 2020; Phelps & Woolford 2021a; Burmeister et al. 2023; Shin et al. 2023). Even when models push probability estimates towards 0.5, they cannot be expected to err in a perfectly sigmoidal fashion. Consequently, it might still be better to use Platt's scaling with a logistic GAM, especially because the GAM will converge to the logistic regression model if the assumptions of logistic regression are met.

If a practitioner cannot justify using Platt's scaling had they not used undersampling, then we recommend either beta calibration or the modification of Platt's scaling with the logit transformation and the GAM. Both approaches are supported analytically when the base model is perfect, but also provide additional flexibility. The GAM, in particular, can fit any smooth relationship if given enough trainining data and performed well when given 1 000 000 observations to learn from. If lots of data is available for training a calibration model, then we recommend this method, as its flexiblity may be valuable in practice when errors in the base model might be more complex than those considered in this study (e.g., asymmetrical about 0.5).

References

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, pp. 641–647, 1955.

R. Benedetti. Scoring rules for forecast verification. *Monthly Weather Review*, 138(1):203–211, 2010.

N. Burmeister, R. D. Frederiksen, E. Høg, and P. Nielsen. Exploration of production data for predictive maintenance of industrial equipment: A case study. *IEEE Access*, 11:102025–102037, 2023.

B. Böken. On the appropriateness of platt scaling in classifier calibration. *Information Systems*, 95:101641, 2021.

K. Coussement and W. Buckinx. A probability-mapping algorithm for calibrating the posterior probabilities: A direct marketing application. *European Journal of Operational Research*, 214(3):732–738, 2011.

A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi. Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence*, pp. 159–166. IEEE, 2015b.

J. de Haan-Ward, S. J. Bonner, and D. Woolford. On the prediction of rare events when sampling from large data. *Communications in Statistics–Simulation and Computation*, pp. 1–21, 2024.

C. Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates, 2001.

T. Guilbert, O. Caelen, A. Chirita, and M. Saerens. Calibration methods in imbalanced binary classification. *Annals of Mathematics and Artificial Intelligence*, pp. 1–34, 2024.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

M. Kull, T. Silva Filho, and P. Flach. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, pp. 623–631. PMLR, 2017.

Meelis Kull, Telmo Silva Filho, and Peter Flach. betacal: Beta calibration for binary classifiers. `https://CRAN.R-project.org/package=betacal`, 2021. R package version 1.0.1.

T. Leathart, E. Frank, G. Holmes, and B. Pfahringer. Probability calibration trees. In *Asian Conference on Machine Learning*, pp. 145–160. PMLR, 2017.

B. Lucena. Spline-based probability calibration, 2018.

E. C. Merkle and M. Steyvers. Choosing a strictly proper scoring rule. *Decision Analysis*, 10(4):292–304, 2013.

J. T. Moreau, T. C. Hankinson, S. Baillet, and R. W. Dudley. Individual-patient prediction of meningioma malignancy and survival using the surveillance, epidemiology, and end results database. *NPJ Digital Medicine*, 3(1):12, 2020.

M. P. Naeini, G. F. Cooper, and M. Hauskrecht. Binary classifier calibration using a bayesian non-parametric approach. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 208–216. SIAM, 2015.

A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005.

F. M. Ojeda, M. L. Jansen, A. Thiéry, S. Blankenberg, C. Weimar, M. Schmid, and A. Ziegler. Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Statistics in Medicine*, 42(29):5451–5478, 2023.

Y. Peng, C. Li, K. Wang, Z. Gao, and R. Yu. Examining imbalanced classification algorithms in predicting real-time traffic crash risk. *Accident Analysis & Prevention*, 144:105610, 2020.

N. Phelps and D. G. Woolford. Comparing calibrated statistical and machine learning methods for wildland fire occurrence prediction: A case study of human-caused fires in Lac La Biche, Alberta, Canada. *International Journal of Wildland Fire*, 30(11):850–870, 2021a.

N. Phelps and D. G. Woolford. Guidelines for effective evaluation and comparison of wildland fire occurrence prediction models. *International Journal of Wildland Fire*, 30(4):225–240, 2021b.

J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, volume 10, pp. 61–74. 1999.

R Core Team. R: A language and environment for statistical computing, 2023.

M. Saerens, P. Latinne, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

H. Shin, S. Shim, and S. Oh. Machine learning-based predictive model for prevention of metabolic syndrome. *PLOS ONE*, 18(6):e0286635, 2023.

S. W. Taylor, D. G. Woolford, C. B. Dean, and D. L. Martell. Wildfire prediction to inform fire management: Statistical science challenges. *Statistical Science*, 28(4):586–615, 2013.

B. M. Turner, M. Steyvers, E. C. Merkle, D. V. Budescu, and T. S. Wallsten. Forecast aggregation via recalibration. *Machine Learning*, 95:261–289, 2014.

D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla. Credit card fraud detection-machine learning methods. In *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, pp. 1–5. IEEE, 2019.

B. C. Wallace and I. J. Dahabreh. Improving class probability estimates for imbalanced data. *Knowledge and Information Systems*, 41(1):33–52, 2014.

S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1): 3–36, 2011.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

A Appendices

A.1 Appendix 1: Simulation study set-up details

A.1.1 Simulating data

For all of the simulated datasets, we generated 10 covariates. Each of these covariates followed a uniform distribution, but with different minimums and maximums. Those values are shown in Table 1.

Table 1: Minimum and maximum values for each of the 10 covariates in the simulated datasets

Covariate	**Minimum**	**Maximum**
1	-0.4	0.6
2	-0.2	0.8
3	-0.4	1.0
4	-0.1	0.9
5	0.0	5.0
6	0.0	3.0
7	1.0	4.0
8	1.0	7.0
9	1.0	3.0
10	0.0	2.0

Based on the 10 covariates, we generated the log odds of success for each observation according to Eq. 10:

$$\begin{aligned}\operatorname{logit}(p) = \frac{\log(99)}{40}\big(&x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} \\ &+ x_1x_3 + x_2x_5 + x_4x_9 + x_6x_7 + x_8x_{10} + x_1x_2x_3x_4 + x_1x_2x_9x_{10}\big) - w\log(99)\end{aligned} \tag{10}$$

Here, w is a parameter that can be used to alter the rate at which successes occur. For our simulation study, we set w to 2, 1.5, and 1.1. Undoing the logit operation yields the success probabilities, which were used to simulate the outcomes in the datasets.

A.1.2 Simulating model outputs

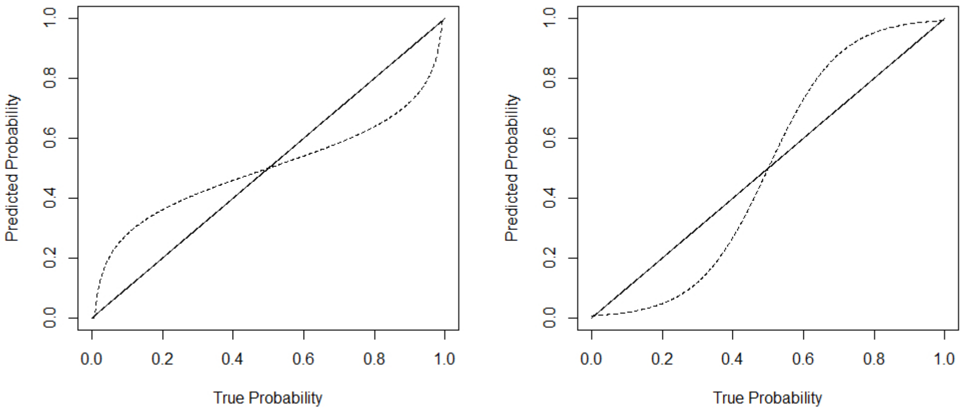

Figure 7: The relationship between the predicted probability and the true probability for a perfect model (solid) and a model with systematic estimation error (dashed). The left plot shows a model that pushes probabilities towards 0.5 (Eq. 8), while the right plot shows a model that pushes probabilities towards 0 and 1 (Eq. 9).

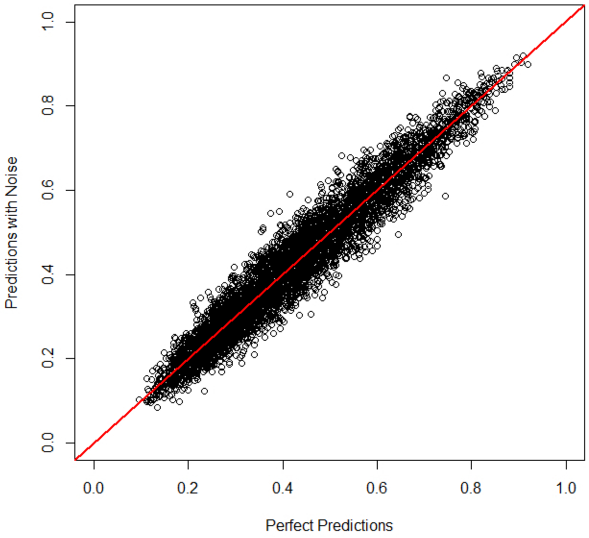

Figure 8: An example of the relationship between the $\hat{\gamma}$'s and the γ's when the base model's predictions are altered by a noise variable. The scatterplot shows 5000 observations obtained from the data generating process with a mean outcome probability of 0.0208. The red line is the 45° line.

A.2 Appendix 2: Results for perfect base model

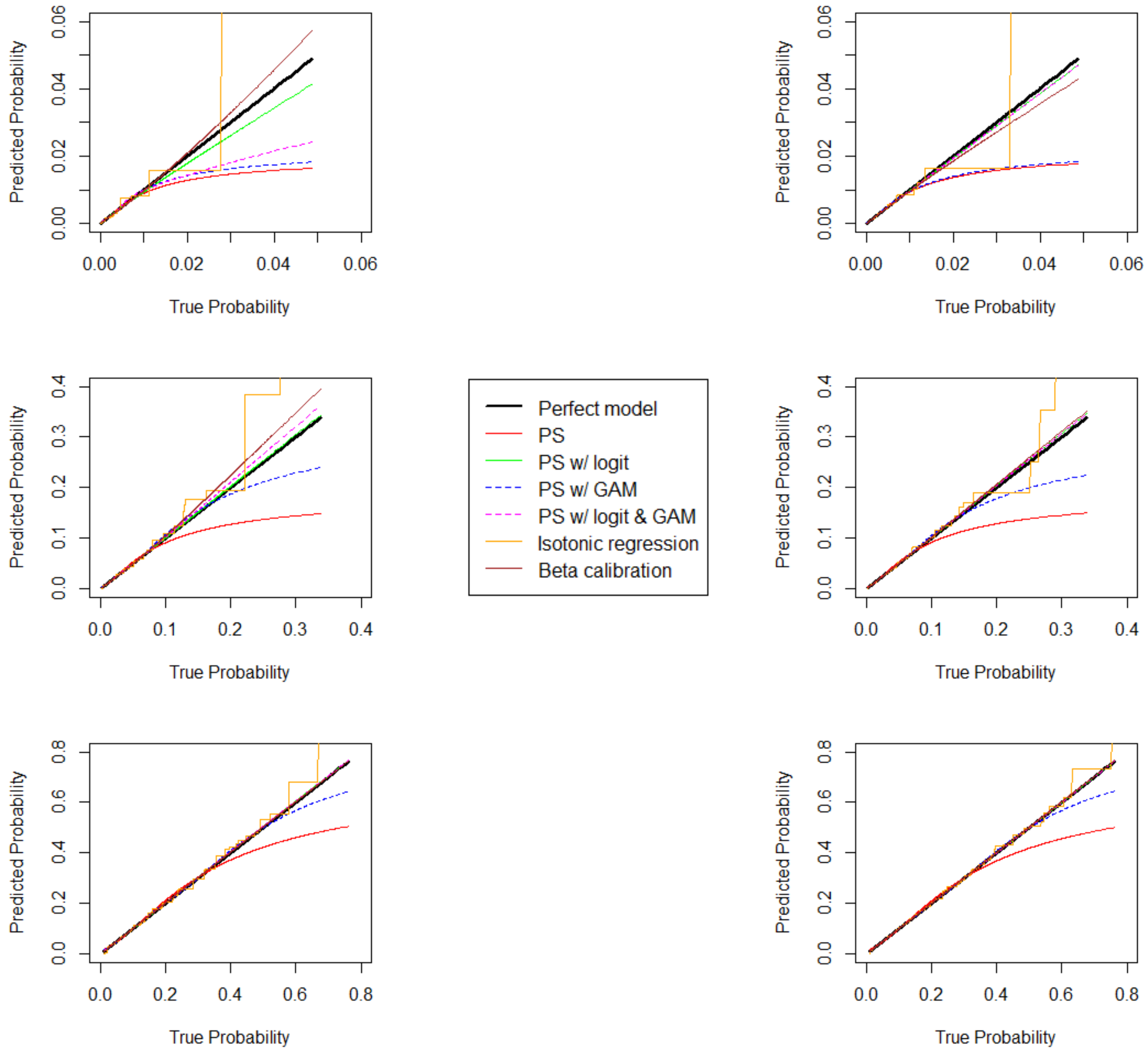

Figure 9: Probability estimates on the testing dataset from Platt's scaling (PS) and its variations, as well as isotonic regression and beta calibration. The data generating processes have a mean probability of success of 0.0022 (top), 0.0208 (middle), and 0.1109 (bottom). The base model is perfect and the calibration methods were trained on 100 000 observations (left) and 1 000 000 observations (right).

Table 2: Root mean squared error (RMSE) and mean absolute error (MAE) for Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration, using 100 000 training observations. The base model used here is perfect.

Success rate	**Calibration method**	**RMSE** ($\times 10^4$)	**MAE** ($\times 10^4$)
0.0022	PS	4.48	1.12
	PS w/ logit	1.77	0.94
	PS w/ GAM	4.49	2.69
	PS w/ logit and GAM	4.50	2.82
	Isotonic regression	8.16	4.24
	Beta calibration	1.42	1.12
	Analytical	0.00	0.00
0.0208	PS	31.09	10.83
	PS w/ logit	2.05	1.49
	PS w/ GAM	12.25	5.37
	PS w/ logit and GAM	5.80	2.40
	Isotonic regression	42.49	13.58
	Beta calibration	12.12	5.24
	Analytical	0.00	0.00
0.1109	PS	87.78	43.88
	PS w/ logit	16.97	13.26
	PS w/ GAM	25.76	14.93
	PS w/ logit and GAM	16.97	13.26
	Isotonic regression	65.47	41.76
	Beta calibration	16.84	13.27
	Analytical	0.00	0.00

Table 3: Root mean squared error (RMSE) and mean absolute error (MAE) for Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration, using 1 000 000 training observations. The base model used here is perfect.

Success rate	**Calibration method**	**RMSE** ($\times 10^4$)	**MAE** ($\times 10^4$)
0.0022	PS	3.91	1.07
	PS w/ logit	0.57	0.26
	PS w/ GAM	3.68	1.00
	PS w/ logit and GAM	0.58	0.26
	Isotonic regression	4.10	1.64
	Beta calibration	1.01	0.26
	Analytical	0.00	0.00
0.0208	PS	30.76	11.22
	PS w/ logit	3.41	1.57
	PS w/ GAM	10.84	3.46
	PS w/ logit and GAM	3.41	1.57
	Isotonic regression	15.25	6.64
	Beta calibration	4.16	1.51
	Analytical	0.00	0.00
0.1109	PS	87.00	40.69
	PS w/ logit	5.91	5.06
	PS w/ GAM	16.34	6.67
	PS w/ logit and GAM	5.91	5.06
	Isotonic regression	34.37	20.56
	Beta calibration	6.31	5.06
	Analytical	0.00	0.00

A.3 Appendix 3: Results for base model pushing probability estimates towards 0.5

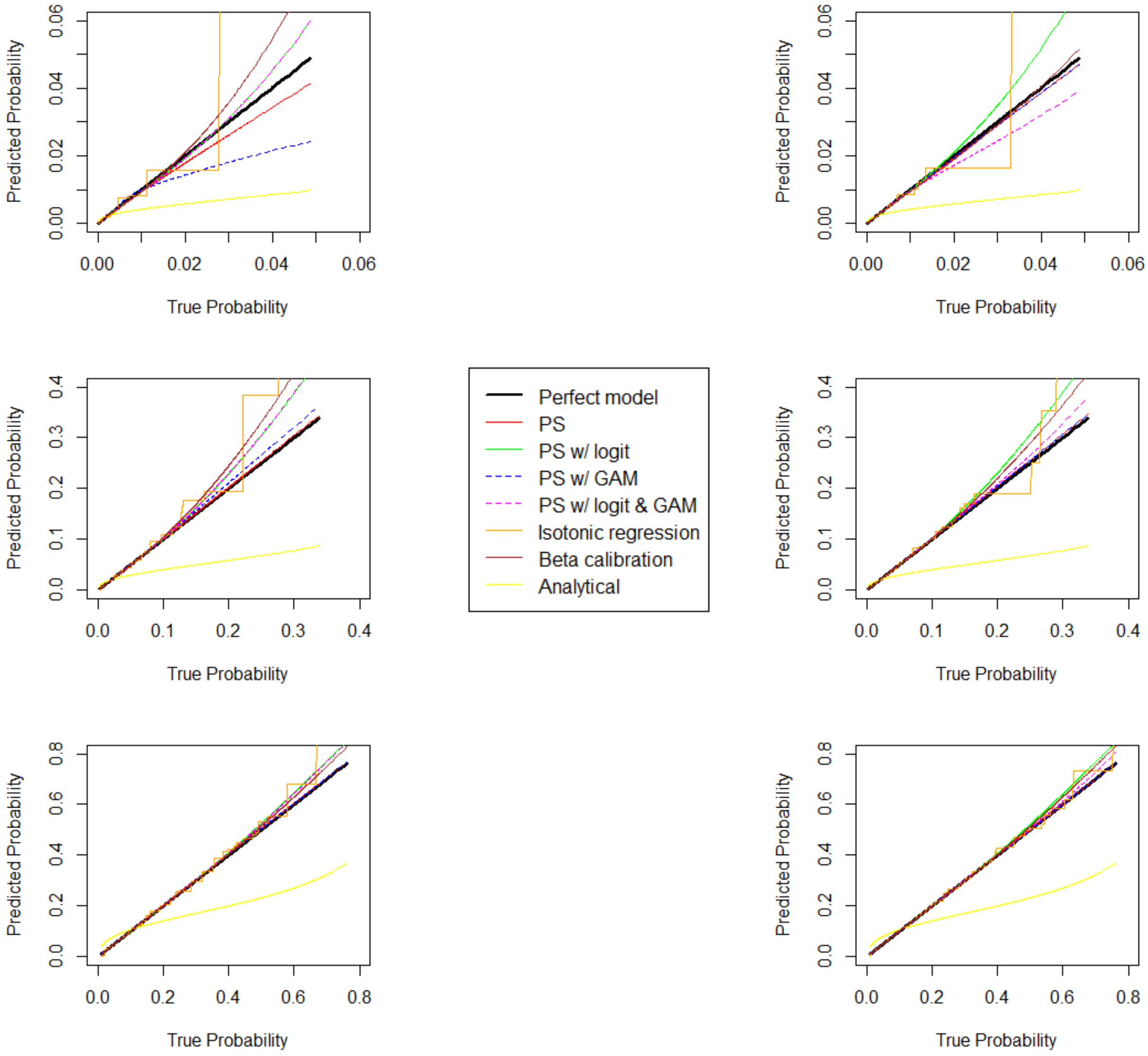

Figure 10: Probability estimates on the testing dataset from Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration. The data generating processes have a mean probability of success of 0.0022 (top), 0.0208 (middle), and 0.1109 (bottom). The base model pushes probability estimates towards 0.5 and the calibration methods were trained on 100 000 observations (left) and 1 000 000 observations (right).

Table 4: Root mean squared error (RMSE) and mean absolute error (MAE) for Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration, using 100 000 training observations. The base model used here pushes probability estimates towards 0.5.

Success rate	**Calibration method**	**RMSE** ($\times 10^4$)	**MAE** ($\times 10^4$)
0.0022	PS	1.77	0.94
	PS w/ logit	1.59	1.08
	PS w/ GAM	4.50	2.82
	PS w/ logit and GAM	1.59	1.08
	Isotonic regression	8.16	4.24
	Beta calibration	1.88	1.19
	Analytical	14.51	8.18
0.0208	PS	2.05	1.49
	PS w/ logit	11.26	2.51
	PS w/ GAM	5.80	2.40
	PS w/ logit and GAM	11.53	2.59
	Isotonic regression	42.49	13.58
	Beta calibration	17.74	4.82
	Analytical	130.10	75.24
0.1109	PS	16.97	13.26
	PS w/ logit	27.15	13.39
	PS w/ GAM	16.97	13.26
	PS w/ logit and GAM	25.47	13.49
	Isotonic regression	65.47	41.76
	Beta calibration	20.50	14.80
	Analytical	506.98	346.29

Table 5: Root mean squared error (RMSE) and mean absolute error (MAE) for Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration, using 1 000 000 training observations. The base model used here pushes probability estimates towards 0.5.

Success rate	**Calibration method**	**RMSE** ($\times 10^4$)	**MAE** ($\times 10^4$)
0.0022	PS	0.57	0.26
	PS w/ logit	1.24	0.44
	PS w/ GAM	0.58	0.26
	PS w/ logit and GAM	1.69	0.46
	Isotonic regression	4.10	1.64
	Beta calibration	1.00	0.41
	Analytical	14.51	8.18
0.0208	PS	3.41	1.57
	PS w/ logit	11.99	2.01
	PS w/ GAM	3.41	1.57
	PS w/ logit and GAM	5.62	2.04
	Isotonic regression	15.25	6.64
	Beta calibration	7.89	1.56
	Analytical	130.10	75.24
0.1109	PS	5.91	5.06
	PS w/ logit	21.77	9.06
	PS w/ GAM	5.91	5.06
	PS w/ logit and GAM	8.03	5.30
	Isotonic regression	34.37	20.56
	Beta calibration	15.55	8.29
	Analytical	506.98	346.29

A.4 Appendix 4: Results for base model pushing probability estimates towards 0 or 1

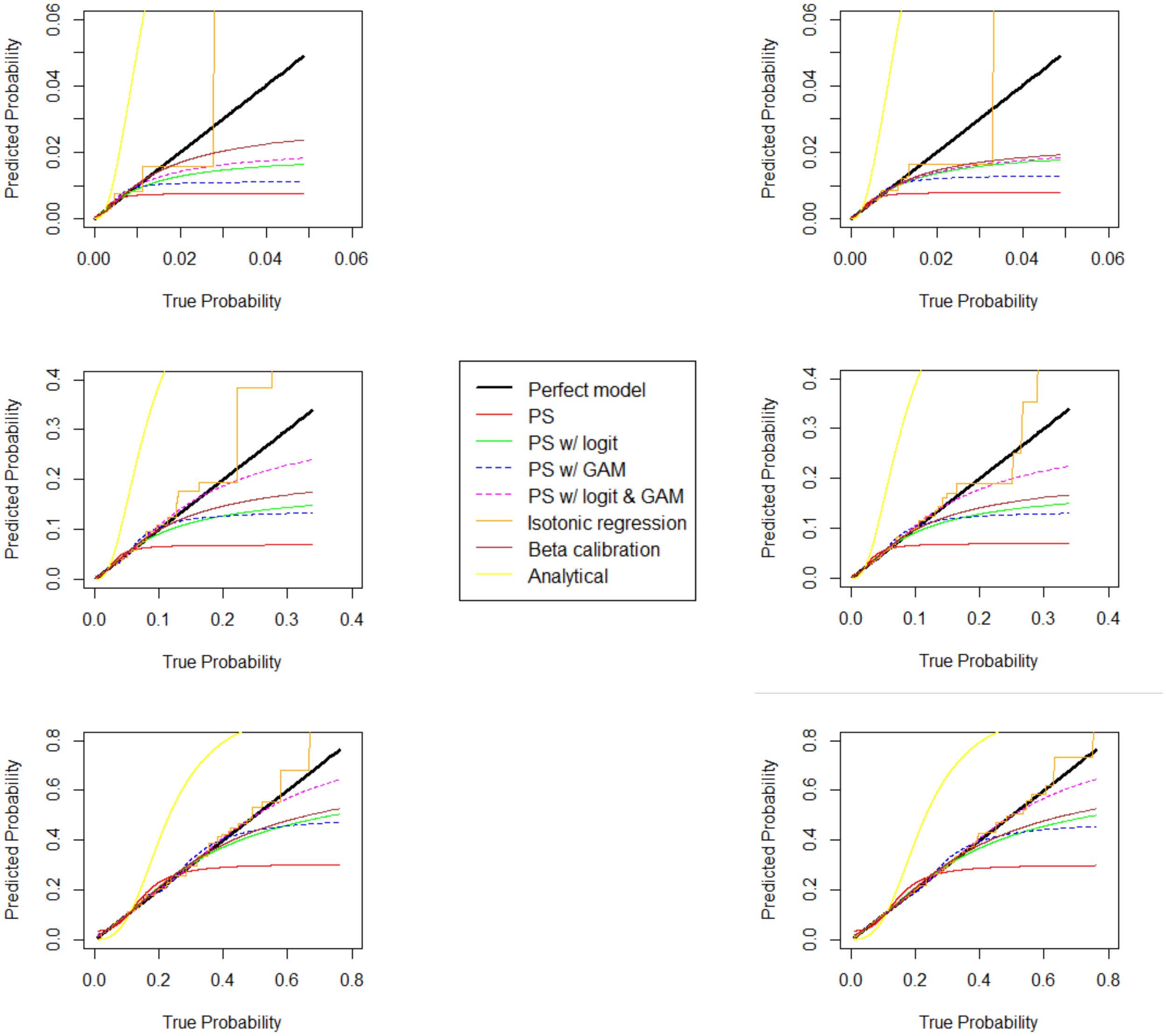

Figure 11: Probability estimates on the testing dataset from Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration. The data generating processes have a mean probability of success of 0.0022 (top), 0.0208 (middle), and 0.1109 (bottom). The base model pushes probabiliy estimates towards 0 or 1 and the calibration methods were trained on 100 000 observations (left) and 1 000 000 observations (right).

Table 6: Root mean squared error (RMSE) and mean absolute error (MAE) for Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration, using 100 000 training observations. The base model used here pushes probability estimates towards 0 or 1.

Success rate	**Calibration method**	**RMSE** ($\times 10^4$)	**MAE** ($\times 10^4$)
0.0022	PS	8.82	3.27
	PS w/ logit	4.48	1.12
	PS w/ GAM	6.81	3.64
	PS w/ logit and GAM	4.49	2.69
	Isotonic regression	8.16	4.24
	Beta calibration	2.77	1.72
	Analytical	88.13	29.67
0.0208	PS	75.23	32.16
	PS w/ logit	31.09	10.83
	PS w/ GAM	32.42	11.74
	PS w/ logit and GAM	12.25	5.37
	Isotonic regression	42.49	13.58
	Beta calibration	22.89	11.34
	Analytical	569.12	238.58
0.1109	PS	248.33	137.80
	PS w/ logit	87.78	43.88
	PS w/ GAM	84.93	41.13
	PS w/ logit and GAM	25.76	14.93
	Isotonic regression	65.47	41.76
	Beta calibration	75.31	40.00
	Analytical	1210.52	735.66

Table 7: Root mean squared error (RMSE) and mean absolute error (MAE) for Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration, using 1 000 000 training observations. The base model used here pushes probability estimates towards 0 or 1.

Success rate	**Calibration method**	**RMSE** ($\times 10^4$)	**MAE** ($\times 10^4$)
0.0022	PS	8.60	3.44
	PS w/ logit	3.91	1.07
	PS w/ GAM	4.75	1.35
	PS w/ logit and GAM	3.68	1.00
	Isotonic regression	4.10	1.64
	Beta calibration	3.37	1.01
	Analytical	88.13	29.67
0.0208	PS	75.01	32.61
	PS w/ logit	30.76	11.22
	PS w/ GAM	32.23	9.88
	PS w/ logit and GAM	10.84	3.46
	Isotonic regression	15.25	6.64
	Beta calibration	24.81	10.35
	Analytical	569.12	238.58
0.1109	PS	248.37	134.63
	PS w/ logit	87.00	40.69
	PS w/ GAM	84.33	33.53
	PS w/ logit and GAM	16.34	6.67
	Isotonic regression	34.37	20.56
	Beta calibration	71.92	37.90
	Analytical	1210.52	735.66

A.5 Appendix 5: Results for base model that is nearly perfect in expectation

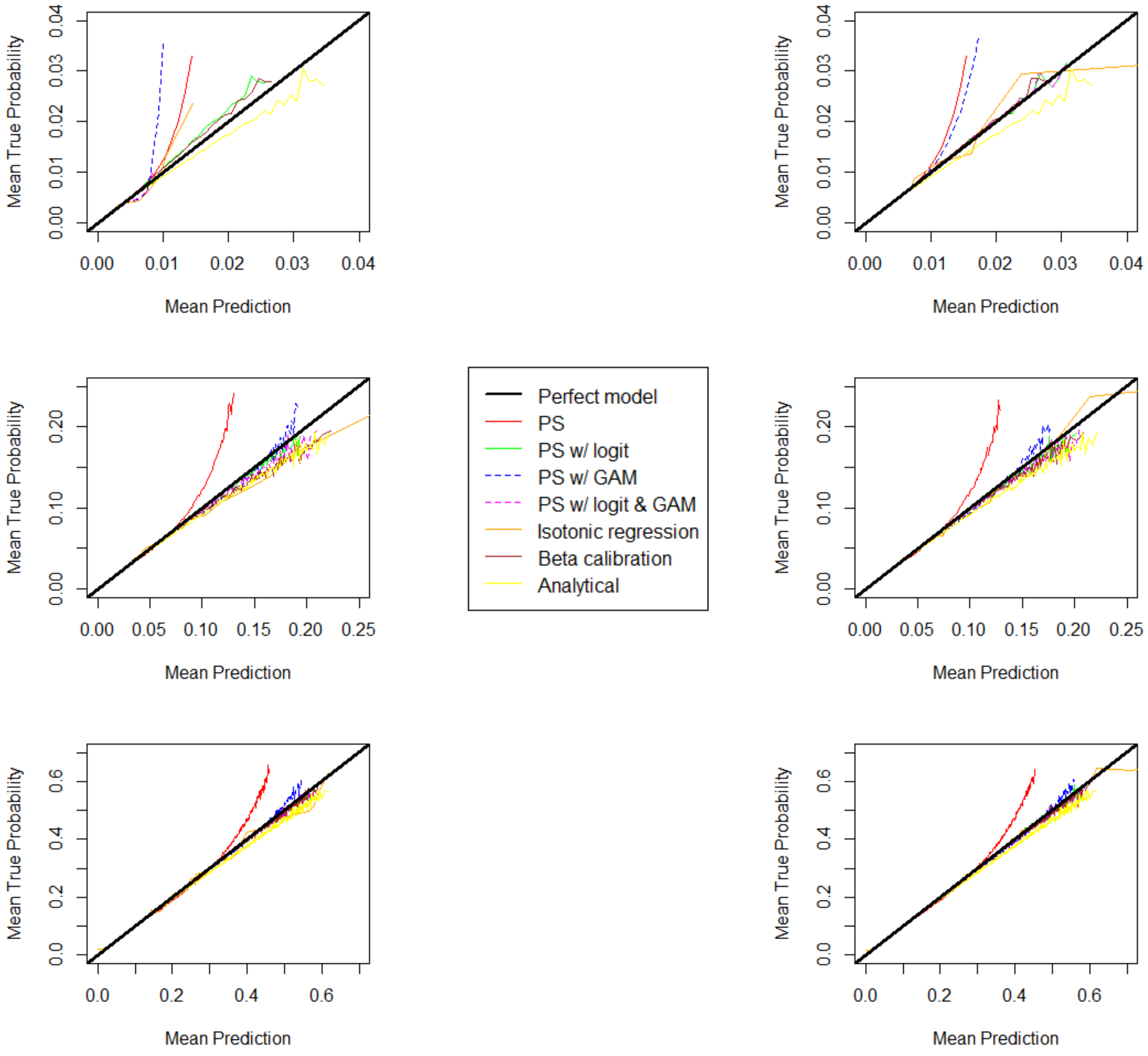

Figure 12: Reliability plots for Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration. The data generating processes have a mean probability of success of 0.0022 (top), 0.0208 (middle), and 0.1109 (bottom). The base model is nearly perfect in expectation and the calibration methods were trained on 100 000 observations (left) and 1 000 000 observations (right).

Table 8: Root mean squared error (RMSE) and mean absolute error (MAE) for Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration, using 100 000 training observations. The base model used here is nearly perfect in expectation.

Success rate	**Calibration method**	**RMSE** ($\times 10^4$)	**MAE** ($\times 10^4$)
0.0022	PS	7.16	3.64
	PS w/ logit	6.08	3.53
	PS w/ GAM	9.94	5.33
	PS w/ logit and GAM	9.62	5.02
	Isotonic regression	9.97	5.82
	Beta calibration	6.02	3.53
	Analytical	6.20	3.50
0.0208	PS	58.88	33.05
	PS w/ logit	50.95	31.08
	PS w/ GAM	52.34	31.48
	PS w/ logit and GAM	51.40	31.12
	Isotonic regression	64.23	34.61
	Beta calibration	52.54	31.39
	Analytical	54.06	32.39
0.1109	PS	223.06	151.10
	PS w/ logit	205.99	144.19
	PS w/ GAM	207.14	144.72
	PS w/ logit and GAM	205.99	144.18
	Isotonic regression	215.20	150.11
	Beta calibration	205.99	144.16
	Analytical	213.50	147.77

Table 9: Root mean squared error (RMSE) and mean absolute error (MAE) for Platt's scaling (PS) and its variations, as well as isotonic regression, beta calibration, and analytical calibration, using 1 000 000 training observations. The base model used here is nearly perfect in expectation.

Success rate	**Calibration method**	**RMSE** ($\times 10^4$)	**MAE** ($\times 10^4$)
0.0022	PS	6.84	3.54
	PS w/ logit	5.79	3.34
	PS w/ GAM	6.57	3.49
	PS w/ logit and GAM	5.79	3.34
	Isotonic regression	6.89	3.72
	Beta calibration	5.80	3.34
	Analytical	6.20	3.50
0.0208	PS	58.69	33.21
	PS w/ logit	51.04	31.20
	PS w/ GAM	51.93	31.36
	PS w/ logit and GAM	51.16	31.19
	Isotonic regression	53.55	32.11
	Beta calibration	51.35	31.20
	Analytical	54.06	32.39
0.1109	PS	222.87	150.24
	PS w/ logit	205.43	143.47
	PS w/ GAM	205.93	143.51
	PS w/ logit and GAM	205.40	143.38
	Isotonic regression	207.59	144.62
	Beta calibration	205.47	143.39
	Analytical	213.50	147.77