Data Science in Finance and Economics

https://www.aimspress.com/journal/dsfe

DSFE, 4(3): 446–468.
DOI:10.3934/DSFE.2024019
Received: 09 April 2024
Revised: 05 August 2024
Accepted: 04 September 2024
Published: 19 September 2024

Research article

Interpretability of the random forest model under class imbalance

Lindani Dube[1,2,*] **and Tanja Verster**[1,2]

[1] Centre for Business Mathematics & Informatics, North West University, Potchefstroom, 2531, Republic of South Africa

[2] National Institute for Theoretical and Computational Sciences (NITheCS), Stellenbosch, 7613, Republic of South Africa

* **Correspondence:** Email: Lindani.Dube@nwu.ac.za.

Abstract: In predictive modeling, addressing class imbalance is a critical concern, particularly in applications where certain classes are disproportionately represented. This study delved into the implications of class imbalance on the interpretability of the random forest models. Class imbalance is a common challenge in machine learning, particularly in domains where certain classes are underrepresented. This study investigated the impact of class imbalance on random forest model performance in churn and fraud detection scenarios. We trained and evaluated random forest models on churn datasets with class imbalances ranging from 20% to 50% and fraud datasets with imbalances from 1% to 15%. The results revealed consistent improvements in the precision, recall, F1-score, and accuracy as class imbalance decreases, indicating that models become more precise and accurate in identifying rare events with balanced datasets. Additionally, we employed interpretability techniques such as Shapley values, partial dependence plots (PDPs), and breakdown plots to elucidate the effect of class imbalance on model interpretability. Shapley values showed varying feature importance across different class distributions, with a general decrease as datasets became more balanced. PDPs illustrated a consistent upward trend in estimated values as datasets approached balance, indicating consistent relationships between input variables and predicted outcomes. Breakdown plots highlighted significant changes in individual predictions as class imbalance varied, underscoring the importance of considering class distribution in interpreting model outputs. These findings contribute to our understanding of the complex interplay between class balance, model performance, and interpretability, offering insights for developing more robust and reliable predictive models in real-world applications.

Keywords: credit; fraud; modeling; classification; imbalance; random forest; interpretability

JEL Codes: C41, C44, C61, G21, G32, O33

1. Introduction

In the ever-evolving landscape of data science and predictive analytics, one of the most pervasive challenges is the intricacy posed by class imbalance within datasets. As organizations delve deeper into leveraging machine learning algorithms to gather insights and drive informed decision-making, understanding how varying class distributions impact model performance becomes paramount.

Consider a decision-maker at a bank, trying to keep customers from leaving. The team made a smart prediction tool, using a statistical model like random forest, to find out who might leave. But as they go through all the data, they keep hitting the same problem: class imbalance. This phenomenon, where one class significantly outnumbers the other(s) in a classification problem (Dube and Verster, 2023), can skew model predictions, leading to biased outcomes and suboptimal decision-making. Motivated by this real-world scenario, our study delves into the intricate interplay between class imbalance and predictive modeling performance. We begin an investigation to understand the hidden patterns, using statistical tools and large amounts of data. Our goal is to figure out how different levels of balanced data affect how well prediction models perform, especially looking at the famous random forest method.

Our investigation stands as a testament to the dynamic nature of predictive modeling. It echoes the sentiments of Verster and Fourie (2023) who delved into the future of predictive modeling by considering the influence of machine learning, financial crises, and financial technology. As we unpack the complexities of class imbalance, we contribute to the broader conversation surrounding the evolving landscape of predictive modeling, paving the way for innovative solutions and collaborative efforts between academia and industry partners. We aim to uncover the nuances hidden within the data, shedding light on the intricate relationship between class imbalance and model behavior. Moreover, we delve further into the essence of the random forest model, employing state-of-the-art techniques such as Shapley values and partial dependence plots. These tools help us navigate the intricate paths of the data and understand the black-box effect. With each analysis, we unravel the intricate web of relationships, shedding light on how individual features influence the model's predictions and how these influences shift with changes in class balance. The ML model's interpretability has gained a lot of attention over the past few decades, with researchers such as Du Toit et al. (2023), Nohara et al. (2022), and Ribeiro et al. (2016) applying it successfully in their research. Jafari et al. (2023), Guliyev and Tatoğlu (2021), Dumitrache et al. (2020) and many more have shown how model interpretability can be used in modeling customer churn. As we explored the data, we found interesting patterns and surprising discoveries.

The analysis of churn and fraud datasets has revealed significant insights in existing literature. Notably, prior studies have demonstrated that addressing class imbalance can lead to substantial improvements in model performance, particularly in the context of precision, recall, F1-score, and accuracy. In our previous work (Dube and Verster, 2024), we have shown how machine learning models for default prediction are affected by missing data and class imbalance, further underscoring the importance of dataset balance in predictive modeling. This study is crucial as it delves into the explainability of model predictions across different levels of class imbalance. By investigating Shapley values and feature importance, the study identifies consistent patterns and significant relationships between features and model predictions. Moreover, PDPs and breakdown plots provide a deeper understanding of how class imbalance affects individual predictions and baseline predictions, highlighting the stability of fundamental relationships between input variables and predicted outcomes as datasets approach balance. Overall, these analyses underscore the importance of addressing class

imbalance for enhancing the performance and reliability of predictive models in identifying rare instances.

The structure of our paper unfolds as follows: We commence with the Introduction in Section 1, providing an overview of the research problem and emphasizing its significance. We outline our dataset in Section 2, comprising imbalanced churn and fraud datasets, and describe the random forest classifier in Section 3. We will review related work on class imbalance and interpretability in Section 4, we introduce various interpretability techniques such as Shapley values, PDPs, and breakdown plots. Section 5 discusses the random forest model and how it can be adopted for the adjustment of class weights. We define evaluation metrics in Section 6 and present results in Section 7 indicating improved model performance with decreased class imbalance. Through discussion in Section 8, we explore the practical implications and underline the importance of considering class distribution for robust model interpretation, concluding with insights for developing reliable predictive models in real-world applications in Section 9.

2. Dataset

In this analysis, we employed two datasets. The first one is a churn dataset sourced from Kaggle, encompassing 10,000 observations with 10 predictor variables and a binary (0/1) response variable. Table 1 represents the description of the churn data and Table 2 shows different sample sizes that were used for the analysis. The second dataset is the fraud dataset, also sourced from Kaggle, with 110,106 observations and eight (8) predictors. Table 3 displays the description of the fraud dataset and Table 4 shows different sample sizes. To generate the different samples of varying class balance, a random over-sampling technique as described in Dube and Verster (2023) was adopted on the minority class. Originally, the churn and fraud datasets had 20% churn and 1% fraud rate, respectively. These are indicated by the asterisk signs on the Tables 2 & 4.

3. Random forest classifier

A random forest (RF) is a classifier made up of a set of tree-structured classifiers (Breiman, 2001), $h(\boldsymbol{x}, \Theta_k)$, where $k = 1, 2, \ldots$. Each tree is built from a random vector of parameters, Θ_k, and contributes a single vote to the most popular class for a given input $\boldsymbol{x}$ (subsample) as indicated in Figure 1 below. This ensemble technique generates diverse classifiers through randomization, resulting in efficient classification, similar to bagging or random subspace methods. The algorithm grows numerous decision trees, and to classify a new object, it goes through each tree in the forest, with the final classification determined by the majority vote across all trees.

Each decision tree is constructed by sampling, with replacement, from the original dataset to form a training set (Liaw et al., 2002). At each node, a subset of input variables is randomly chosen for splitting, ensuring diversity among the trees. In our case, a maximum of 2 features were specified in the model when looking for the best split at each node. By setting this parameter to a value less than the total number of features in the dataset, a random subset of features will be considered for splitting at each node. This helps introduce diversity among the trees in the ensemble. The design parameters include the number of features selected for each tree, the number of trees in the forest, and the minimum number of samples in a leaf node. Notably, the selection of features significantly impacts the RF's performance. An important aspect of RF is the use of out-of-bag (OOB) data, which consists of approximately one-third

Table 1. Churn dataset description.

Feature	Description
Customer ID	Unused
Credit score	Input
Country	Input
Gender	Input
Age	Input
Tenure	Input
Balance	Input
Products number	Input
Credit card	Input
Active member	Input
Estimated	Input
Churn	Target

Table 2. Churn dataset samples.

Churn %	Yes	No
20*	2,037	7,963
30	3,583	7,963
40	5,574	7,963
50	7,963	7,963

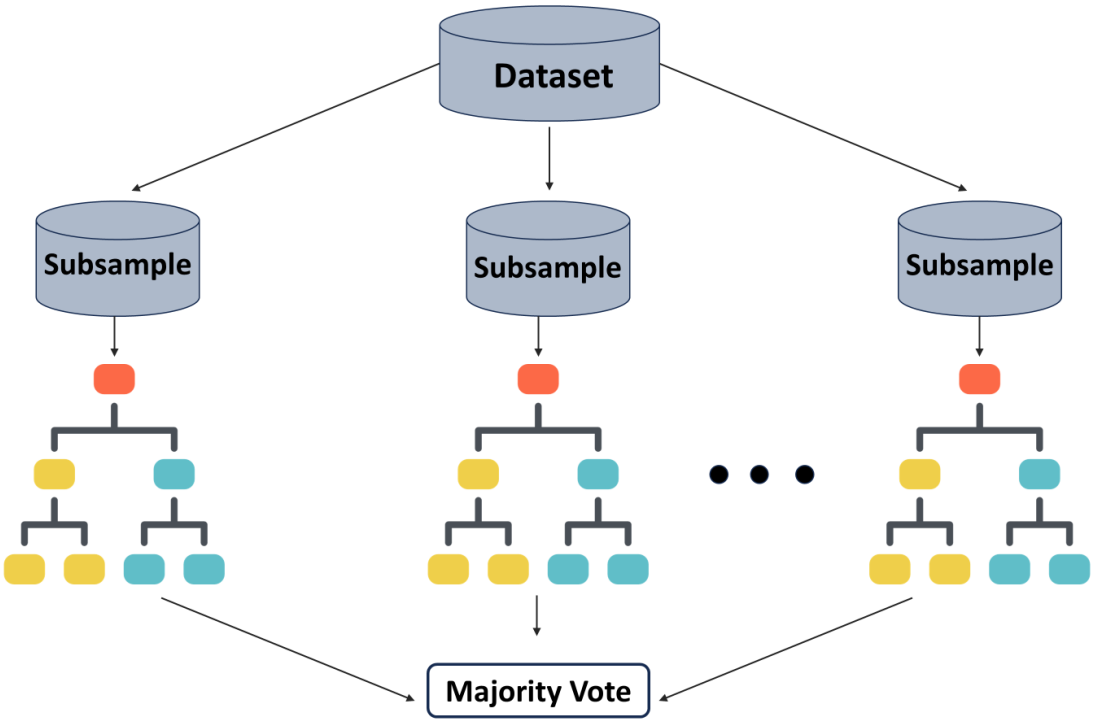

Figure 1. Architecture of a random forest classifier.

Table 3. Fraud dataset description.

Feature	Description
Fraud	Fraud transaction, indicator variable
Type	Type of online transaction
Amount	The amount of the transaction
OldbalanceOrg	Balance before the transaction
NewbalanceOrig	Balance after the transaction
OldbalanceDest	Initial balance of recipient before the transaction
NewbalanceDest	The new balance of recipient after the transaction

Table 4. Fraud dataset samples.

Fraud %	Yes	No
1*	1,059	109,047
5	5,452	109,047
10	10,905	109,047
15	16,357	109,047

of the original dataset not included in the bootstrap sample (Gislason et al., 2006). This OOB data facilitates unbiased estimation of classification error, eliminating the need for separate validation sets or cross-validation. The accuracy of RF is characterized by its generalization error, which is determined by the margin function. This function measures the difference between the average number of votes for the correct class and the maximum average vote for any other class. The strength of RF, in terms of the margin function, reflects its ability to reduce variance through averaging and randomization, thereby decreasing correlation among the trees in the forest (Abd Algani et al., 2022); (Liaw et al., 2002). In this analysis, a subset of only 2 input variables (features) was randomly chosen for splitting, ensuring diversity among the trees and the forest contained 100 decision trees, with each trained on a bootstrap sample of the training data with replacement.

Breiman (2001) highlights several strengths of random forest, including its efficiency on large databases, robustness to datasets with thousands of input variables, estimation of important variables, handling of missing data, and ability to balance class errors in imbalanced datasets. Mathematically, the generalization error of the ensemble classifier is bounded above by a function of the mean correlation between base classifiers and their average strength (Hastie et al., 2009). If ρ represents the mean correlation, the upper bound for the generalization error is given by $\rho(1 - S^2)/S^2$, where S is the expected value of the strength of the random forest.

In our study, we extended the application of RF to address class imbalance, a common challenge in binary classification tasks. As highlighted by Dube and Verster (2023), RF demonstrates superior performance in handling class imbalance compared to other machine learning models. To further enhance the interpretability and effectiveness of the RF model in our analysis, we employed the

technique of RF with class weights (Shahhosseini and Hu, 2021). This approach involves modifying the weighting strategy of the standard RF model, assigning higher weights to the minority class instances during training. By incorporating class weights, the RF model can effectively correct for oversampling and make more accurate predictions, as demonstrated by Winham et al. (2013). Through this adaptation, RF with class weights aims to mitigate the bias toward the majority class and improve the overall balance and performance of the classifier, ensuring fair treatment of both classes in the binary classification setting.

In accordance with the guidelines outlined by Nationalbank Oesterreichische (2004), it is imperative to adjust the probabilities obtained from oversampled samples to align with the average probabilities of the original dataset. This adjustment is achieved indirectly using relative default frequencies (RDFs), as specified in the following procedure:

1. Compute the average sample default rate derived from the random forest model and transform it into $\text{RDFs}_{\text{sample}}$.
2. Determine or estimate the average default rate in the original dataset and convert it into $\text{RDF}_{\text{original}}$.
3. Calculate the representation of each default probability generated by the random forest model as $\text{RDF}_{\text{unscaled}}$.
4. Multiply $\text{RDF}_{\text{unscaled}}$ by the scaling factor specific to the corresponding model.
5. Convert the resulting scaled RDF into a scaled default probability.

The scaled $\text{RDF}_{\text{scaled}}$ is computed as follows:

$$RDF_{scaled} = RDF_{unscaled} \times \frac{RDF_{original}}{RDF_{sample}}$$

Here, RDF denotes the probability of default (PD) divided by 1 − PD or $PD = \frac{RDF}{1+RDF}$. $\text{RDF}_{\text{sample}}$ is derived from the average predicted probability of default within our implementation sample, while $\text{RDF}_{\text{original}}$ reflects the true default rate in the original dataset prior to oversampling. Lastly, $\text{RDF}_{\text{unscaled}}$ is computed from the individual default probabilities generated by the random forest model. This procedure ensures the calibration of PDs to accurately reflect the characteristics of the original dataset while considering the effects of oversampling.

Our methodology (outlined in Figure 2) initiates by acquiring the dataset and meticulously cleaning it to ensure data integrity. Samples of varying class imbalance were generated in order to assess the impact on the performance of an RF model. These samples were then divided into distinct training and testing subsets, facilitating both model training and evaluation. During the training phase, the random forest classifier is trained using the training subset, while the testing subset is reserved for assessing the model's performance. After generating predicted probabilities, we adopted the approach proposed by Nationalbank Oesterreichische (2004) to transform these probabilities, ensuring they accurately reflect the characteristics of the true population. Subsequently, we meticulously reported on the performance measures outlined in Section 6.

4. Related Work

Interpretability in machine learning ensures trustworthiness and comprehension of model decisions, particularly in domains where such decisions carry significant implications. Across various studies, the

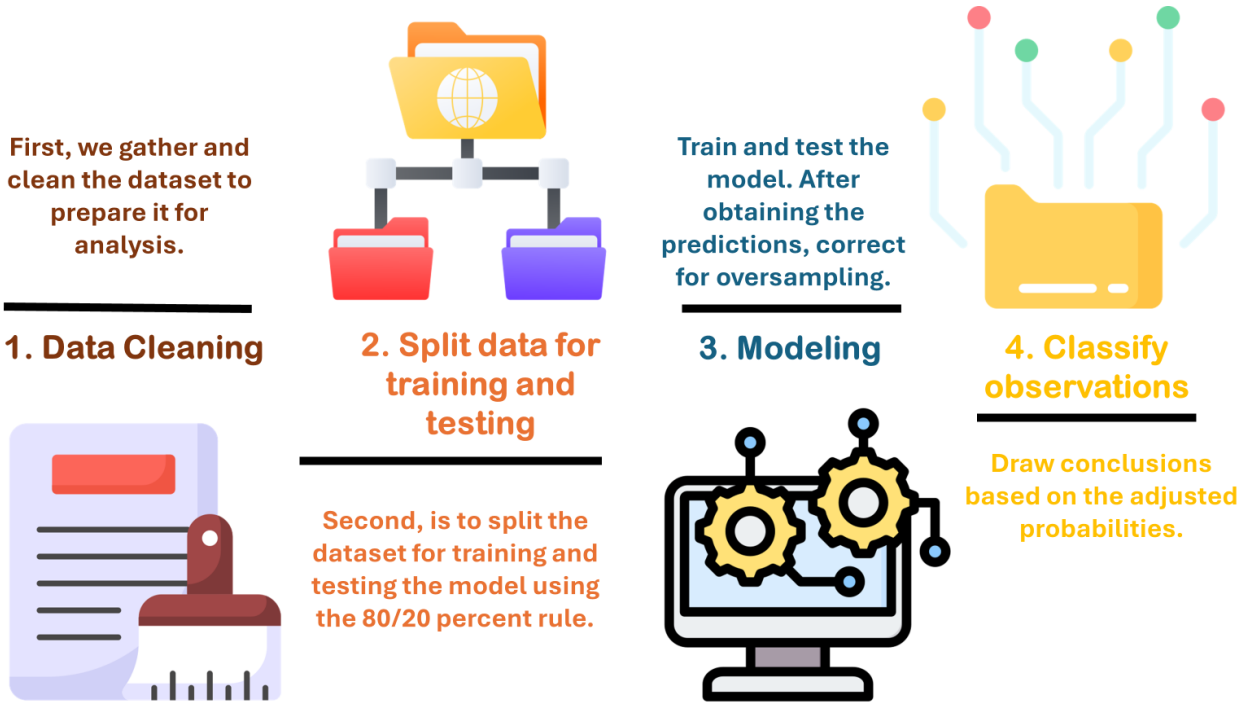

Figure 2. Analysis approach.

importance of interpretability resonates as researchers navigate the complexities of diverse applications.

In the context of customer churn prediction, Jafari et al. (2023) proposed a comprehensive framework aimed at enhancing both predictive performance and interpretability. Their approach, spanning preprocessing techniques, novel classification algorithms, and rigorous evaluation criteria, addresses the dual challenge of accurate prediction and transparent decision-making, catering to the needs of managerial stakeholders. Similarly, Tekouabou et al. (2022) tackled the intricacies of customer relationship management systems, recognizing the challenges posed by heterogeneous data and class imbalances. Through the adept application of ensemble methods and data balancing techniques, they constructed predictive models that not only mitigate these challenges but also offer interpretable insights, facilitating informed decision-making within CRM contexts. In the banking sector, Peng et al. (2023) delved into the pressing issue of customer churn, leveraging advanced modeling techniques augmented by interpretability analyses. By employing genetic algorithm-enhanced XGBoost and elucidating feature contributions through Shapley values, they provided actionable insights for banking institutions, empowering them to proactively address customer retention challenges.

Building upon the insights gleaned from existing research, Zhu et al. (2023) and Davis et al. (2022) offered valuable contributions by employing a range of algorithms such as LightGBM, XGBoost, logistic regression, and decision trees to forecast loan defaults. These models not only exhibited high predictive performance, as evidenced by metrics like accuracy and area under the curve, but also prioritized interpretability through methods like local interpretable model-agnostic explanations (LIME) and generated simple rules understandable to various stakeholders. Similarly, Ariza-Garzón et al. (2020) and Tran et al. (2022) underscored the significance of explainable credit risk models in peer-to-peer lending and financial markets. By utilizing advanced techniques like SHAP values, they demonstrated

how machine learning algorithms can not only achieve superior predictive accuracy but also offer transparency and comprehensibility which was deemed crucial for fostering trust among stakeholders including industry players, regulators, and investors.

In nanoparticle studies, Yu et al. (2021) navigated the complexities of highly heterogeneous data, developing a framework that combines tree-based random forest analysis with feature interaction networks. Their approach not only facilitates accurate prediction of immune responses and lung burden but also enhances model interpretability, thereby offering valuable guidance for nanoparticle design and application. Meanwhile, Uddin et al. (2022) focused on credit default prediction, employing random forest methodology to discern patterns within micro-enterprise credit data. Through rigorous analysis and consideration of both traditional financial variables and non-traditional predictors, they underscore the importance of interpretability in credit risk assessment, offering insights that are invaluable for financial market participants. Lastly, Moraffah et al. (2020) provided a comprehensive survey on causal interpretable models, shedding light on the evolving landscape of interpretability methodologies. By exploring the nuances of causal explanations and evaluation metrics, they equip practitioners with a deeper understanding of interpretability concepts, thereby fostering greater transparency and trust in machine learning systems.

Collectively, these studies and more underscore the critical role of interpretability in enhancing the utility and reliability of machine learning models across diverse domains, offering insights that are indispensable for informed decision-making and stakeholder trust.

5. Machine learning interpretability

This section explores several key methods for understanding model behavior and feature importance. We delve into permutation feature importance, Shapley values, partial dependence plots, and breakdown plots, each providing unique perspectives on model interpretability. Permutation feature importance uncovers the significance of individual features by assessing the impact of shuffling feature values. Shapley values, rooted in cooperative game theory, assign values to features based on their contribution to predictions for specific instances. Partial dependence plots offer insights into the relationship between features and predictions by visualizing how the prediction changes with varying feature values. Finally, breakdown plots provide a granular view of feature contributions to individual predictions, aiding in model debugging and transparency. These techniques collectively enhance our understanding of machine learning models and promote trust, transparency, and fairness in decision-making processes. In the following subsections, we will discuss these interpretability techniques in details.

5.1. Permutation feature importance

Researchers need to identify the primary predictor in a predictive model and ascertain its comparative impact on model outcomes. Permutation importance, employed by Breiman (2001), is a commonly employed method to assess feature significance. It involves randomly shuffling feature values and observing resultant changes in model predictions to discern which features influence predictions most significantly. Importance weights are determined based on the predictive variance between the original and perturbed feature values (Fisher et al., 2019). Feature importance, inferred from these weights, can be evaluated for all features, providing insight into their respective impacts on model outputs (Gregorutti et al., 2017). Permutation importance for features can be expressed as:

$$I(j) = \exp(f(x_{+j})) - \exp(f(x_{+j} + \pi(x_j))). \tag{1}$$

Here, j indicates the j^{th} feature that needs explanation, x_j denotes the value of the j^{th} feature, and x_{+j} indicates the value of sample x with the j^{th} feature. $\pi(x_j)$ denotes the disturbance added to x_j. f is the prediction of a complex model on x and exponential expression (exp()) is the predicted accuracy of f.

5.2. Shapley values

According to Shapley (2020) and Lundberg and Lee (2017), Shapley values are a concept from cooperative game theory. In machine learning, they are used to assign a value to each feature that represents its contribution to the prediction for a specific instance. The concept aims to distribute the total gain or payoff among players based on their relative contributions to the final outcome of a game. Shapley values offer a method to fairly allocate rewards to each player, characterized by natural properties such as local accuracy (additivity), consistency (symmetry), and nonexistence (null effect) (Shapley, 2020). In the context of activity predictions, Shapley values can also be interpreted as a fair allocation of feature importance given a specific model output (Rodríguez-Pérez and Bajorath, 2019). Features contribute differently to the model's output, which is captured by Shapley values, representing both the magnitude and direction of the contribution. Features with positive values contribute to activity prediction, while those with negative values contribute to inactivity prediction.

The importance of a feature j is quantified by its Shapley value, as defined in Equation 2:

$$\phi_j = \frac{1}{|N|!} \sum_{S \subseteq N \setminus \{j\}} |S|!(|N| - |S| - 1)! \left[f(S \cup \{j\}) - f(S) \right] \tag{2}$$

where $f(S)$ is the model output with a feature set S, and N is the complete set of features. The Shapley value of feature j (ϕ_j) is computed as the average of its contributions across all possible permutations of feature sets. This approach accounts for feature orderings, crucial for understanding changes in model output due to correlated features.

5.3. Partial dependence plots

The concept of the partial dependence profile (PDP) was introduced by Greenwell et al. (2017) . Let j denote any j^{th} feature in the dataset. Then, the PDP can be defined as a function of the observation z for a model f and a variable j as follows:

$$PDP(f, j, z) = E_{-j}[f(j^{|=z})]. \tag{3}$$

In simpler terms, the PDP value for the j^{th} column in the observation z is the average prediction of model f when values in the j^{th} column are set to z. However, in practice (Biecek and Burzykowski, 2021a), the distribution of $-j$ is often unknown. Therefore, it is estimated using the following formula:

$$\widehat{PDP}(f, j, z) = \frac{1}{n} \sum_{i=1}^{n} f(j_i^{|=z}). \tag{4}$$

5.4. Breakdown plot

A breakdown (BD) plot (Biecek and Burzykowski, 2021b) shows the contributions of each feature to the final prediction for a single instance. It visually breaks down the prediction into the impact of individual features. This approach offers a model-agnostic method for interpreting predictions, allowing for the explanation of both additive and non-additive models. While it may lead to some loss of information regarding the model's structure, it proves useful for various models. The core idea behind the ag-break approach is to identify elements of x_{new} that, if altered significantly, would result in a notable change in the prediction $f(x_{\text{new}})$. This approach uses the concept of a relaxed model prediction (Staniak and Biecek, 2018). Let $f_{\text{IndSet}}(x_{\text{new}})$ denote the expected model prediction for x_{new} relaxed on the set of indices IndSet $= \{1, \ldots, p\}$.

$$f^{\text{IndSet}}(x^{\text{new}}) = E[f(x) \mid x_{\text{IndSet}} = x^{\text{new}}_{\text{IndSet}}].$$

The relaxed prediction represents an average model response for observations matching x_{new} for features in IndSet^C, following the population distribution for features in IndSet.

Since the joint distribution of x is unknown, an estimate is used instead:

$$\widehat{f^{\text{IndSet}}(x^{\text{new}})} = \frac{1}{n}\sum_{i=1}^{n} f(x^{i}_{-\text{IndSet}}, x^{\text{new}}_{\text{IndSet}}).$$

Individual prediction explanations explains why a specific prediction was made and which features had the most influence. In our case, individual explanations will be adopted to help explain the impact of oversampling the minority cases. Particularly, this will explain how individual predictions are affected.

6. Evaluation metrics

In this paper, we adopted a widely used approach to understanding the performance of a random forest model in handling class imbalance, namely precision, recall, and F1-score as outlined by Goutte and Gaussier (2005). These metrics play a critical role in assessing the performance of classification models and are essential for determining their effectiveness in real-world applications.

Accuracy measures the overall correctness of the model's predictions across all classes (Jiao and Du, 2016). It is calculated as the ratio of correctly predicted instances to the total number of instances in the dataset, as shown in Equation 5:

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Instances}. \tag{5}$$

A high accuracy indicates that the model is making correct predictions across all classes. However, accuracy alone may not be sufficient for evaluating the performance of a model, especially in the presence of imbalanced datasets where one class dominates the others.

Precision, also known as positive predictive value, measures the accuracy of positive predictions made by the model. It is calculated as the ratio of true positive predictions to the total number of positive predictions, as shown in Equation 6:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}. \tag{6}$$

A high precision indicates that the model is proficient at correctly identifying positive instances while minimizing false positives.

Recall, also referred to as sensitivity, measures the ability of the model to capture all positive instances in the dataset. It is calculated as the ratio of true positive predictions to the total number of actual positive instances, as shown in Equation 7:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}. \tag{7}$$

A high recall indicates that the model can successfully identify most positive instances, minimizing false negatives.

F1-score is the harmonic mean of precision and recall, providing a balanced assessment of a model's performance. It is calculated using Equation 8:

$$F1\text{-}score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{8}$$

The F1-score considers both false positives and false negatives, making it a useful metric for evaluating models with imbalanced datasets. Precision, recall, and F1-score are essential metrics in machine learning for evaluating the performance of classification models. While precision focuses on the accuracy of positive predictions, recall emphasizes the model's ability to capture all positive instances. The F1-score provides a balanced measure by considering both precision and recall, making it a valuable tool for model evaluation. These measures consider the number of positive and negative cases and to accommodate for the rare cases, we will adopt the methodology specified in Section 3.

7. Results

In the pursuit of understanding the influence of class imbalance on model performance, a random forest model was trained and evaluated on a churn (20%, 30%, 40%, and 50%) and fraud (1%, 5%, 10%, and 15%) dataset with varying levels of class distribution. Both datasets that were used underwent an 80/20 split into training and testing sets. The random forest model was trained on four different samples (per original dataset), each with varying class balance proportions. The subsequent testing results across these different churn and fraud percentages are detailed in Table 5 below. The scores on the table strictly represent the positive cases.

Table 5. Summary of testing results obtained.

Dataset	Class %	Precision	Recall	F1-score	Accuracy
Churn	20*	45	78	57	76
	30	64	84	73	81
	40	76	83	80	82
	50	83	80	82	83
Fraud	1*	17	49	29	94
	5	53	52	52	95
	10	66	54	59	97
	15	69	60	63	98

In the churn dataset analysis, we observed a consistent improvement in precision, recall, F1-score, and accuracy as the class imbalance decreased. Precision, which measures the proportion of true positive predictions among all positive predictions, showed an increase from 45% to 83% as the class imbalance decreased from 20% to 50%. This suggested that with a more balanced dataset, the model becomes more precise in correctly identifying churn cases. Recall, representing the proportion of true positive predictions among all actual positives, also demonstrated improvement from 78% to 80% with decreasing class imbalance. This indicates that the model is better at capturing actual churn cases when the dataset is less imbalanced. F1-score, which is the harmonic mean of precision and recall, showed a similar trend of enhancement from 57% to 82% as class imbalance decreased. This implies that the overall performance of the model in balancing precision and recall improved with a more balanced dataset. Accuracy, reflecting the proportion of correctly classified cases among all cases, increased from 76% to 83% as class imbalance decreased. This indicates that the model's overall predictive accuracy improves with a reduction in class imbalance, as it becomes better at correctly classifying both churn and non-churn cases.

In the fraud dataset analysis, we also observed a consistent improvement in precision, recall, F1-score, and accuracy as the class imbalance decreased. Precision increased from 17% to 69% as the class imbalance decreased from 1% to 15%. This suggests that with a more balanced dataset, the model becomes more precise in identifying fraud cases. Recall showed a significant improvement from 49% to 60% with decreasing class imbalance, indicating that the model captured a higher proportion of actual fraud cases when the dataset was less imbalanced. F1-score demonstrated a similar trend of enhancement from 29% to 63% as class imbalance decreased, implying an overall improvement in the model's ability to balance precision and recall. Accuracy increased from 94% to 98% as class imbalance decreased, indicating an overall improvement in the model's predictive accuracy with a reduction in class imbalance.

The next part of the experiment was to investigate the impact, or rather the effect, class imbalance has on explaining this sophisticated model. First, Shapley values were investigated across the four samples as shown in Figures 3–17. The Figures 3, 5, 7, 9 display Shapley values for each feature and instance in the churn dataset. The vertical position indicates the feature, and the horizontal position shows the Shapley value. The color shows the feature value, ranging from low to high. If points overlap, they are slightly moved vertically to show the spread of Shapley values for each feature. Features are arranged based on their importance. Figures 11, 13, 15, 17 display the same information but with the focus on the feature importance. Noticeably, it was observed that features age and balance had a positive relationship with the Shapley values throughout the four samples whereas variables such as products number, active member, and credit card showed negative relationships with the Shapley values. Some of the features, like country, did not show the same relationship throughout the samples. We also noted the reordering of features from 20%–40% class imbalanced which then stayed the same when the dataset was 50% balanced. Moreover, we observed an overall decrease of Shapley values as the dataset became more balanced but an improvement in feature importance. In the Fraud dataset, the ordering of features in terms of importance was also observed and the overall decrease of the SHAP values in all the samples.

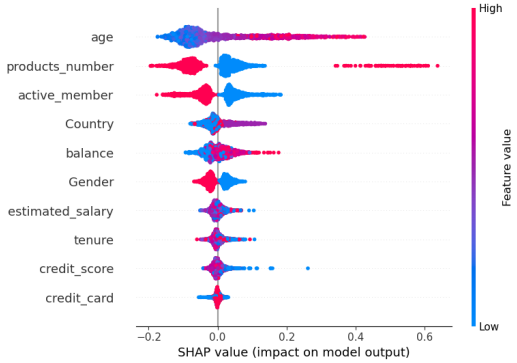

Figure 3. 20% Churn rate.

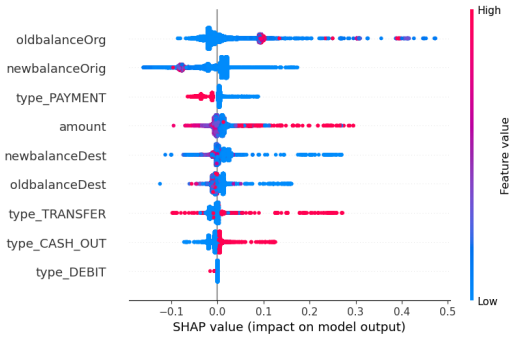

Figure 4. 1% Fraud rate.

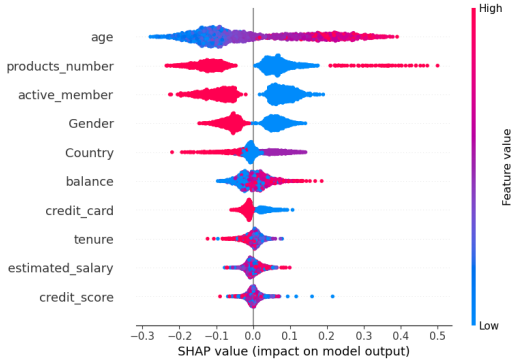

Figure 5. 30% Churn rate.

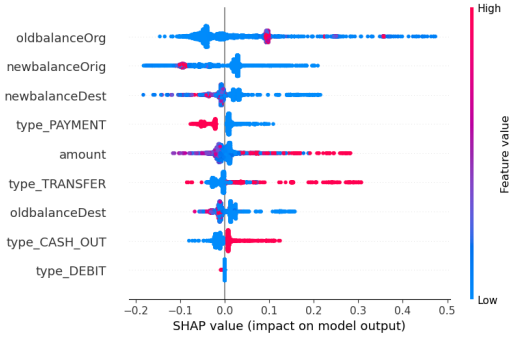

Figure 6. 5% Fraud rate.

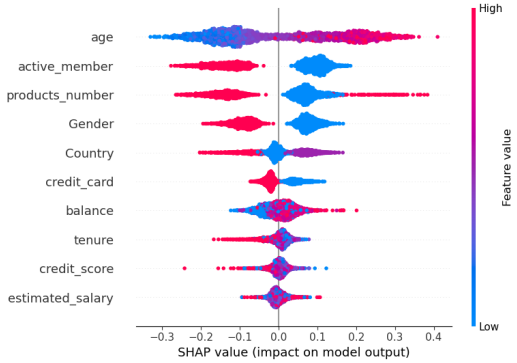

Figure 7. 40% Churn rate.

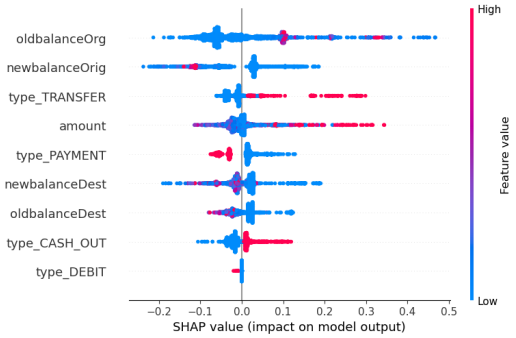

Figure 8. 10% Fraud rate.

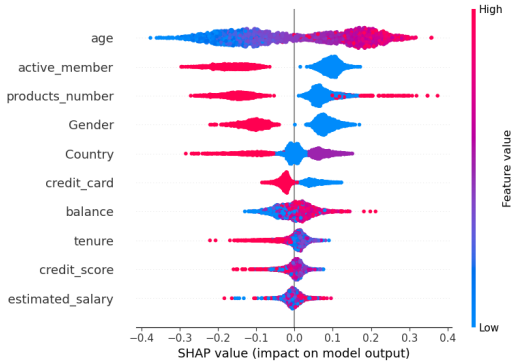

Figure 9. 50% Churn rate.

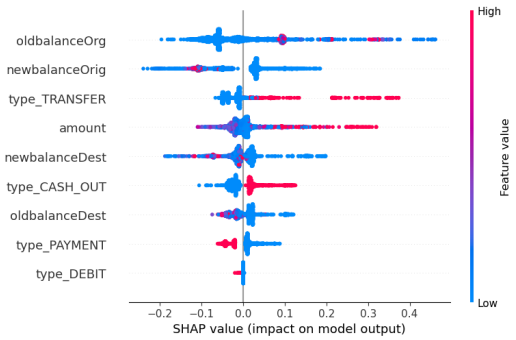

Figure 10. 15% Fraud rate.

Shapley values for churn and fraud datasets

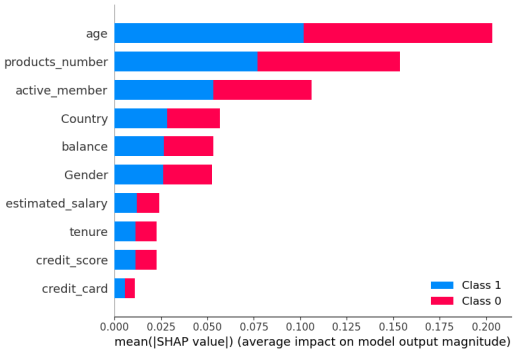

Figure 11. 20% Churn rate.

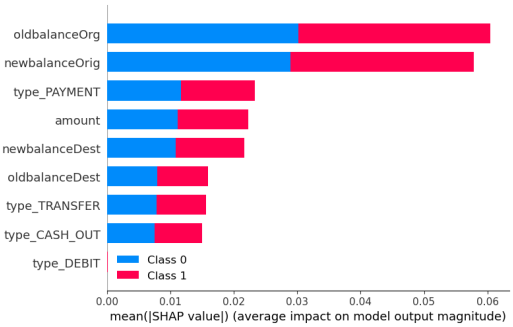

Figure 12. 1% Fraud rate.

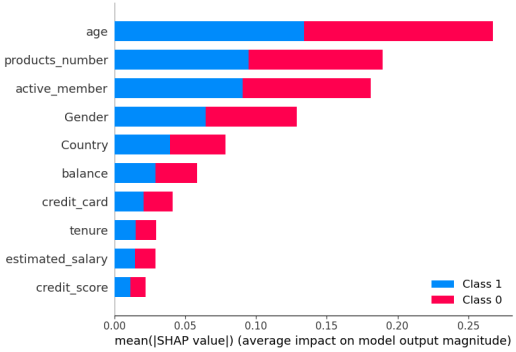

Figure 13. 30% Churn rate.

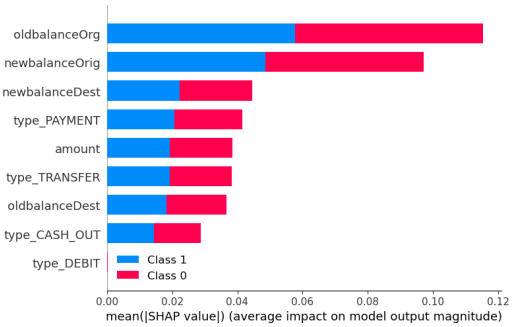

Figure 14. 5% Fraud rate.

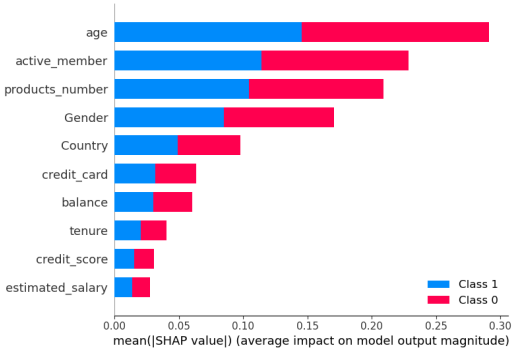

Figure 15. 40% Churn rate.

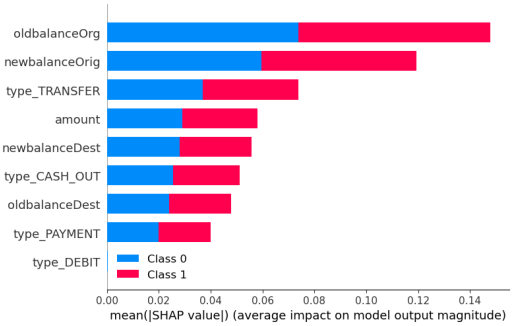

Figure 16. 10% Fraud rate.

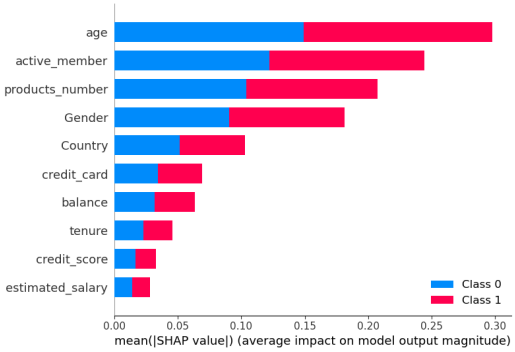

Figure 17. 50% Churn rate.

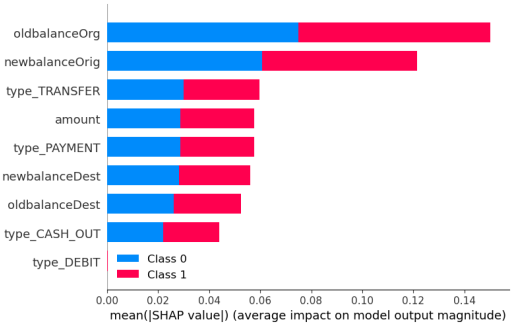

Figure 18. 15% Fraud rate.

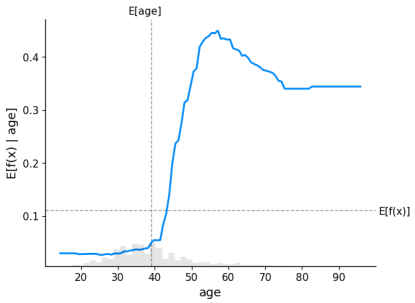

Figure 19. PDP 20%.

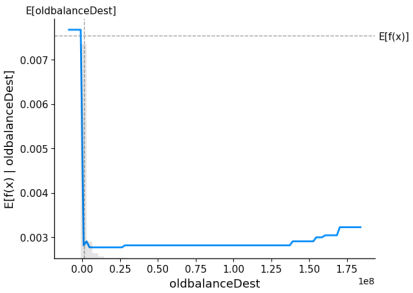

Figure 20. PDP 1%.

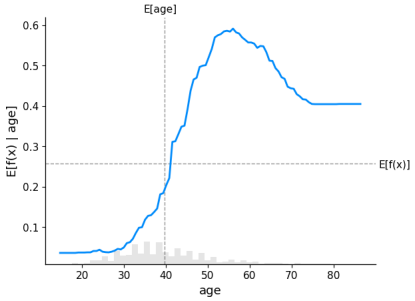

Figure 21. PDP 30%.

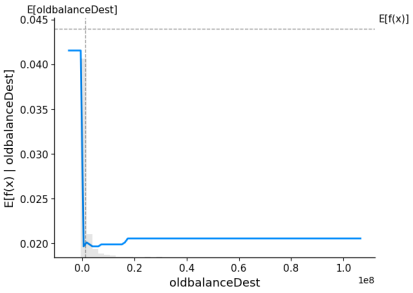

Figure 22. PDP 5%.

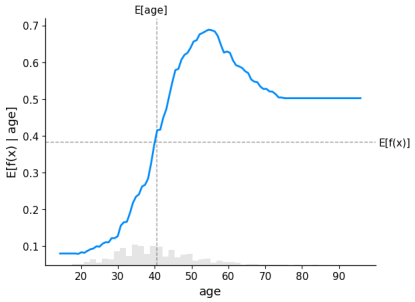

Figure 23. PDP 40%.

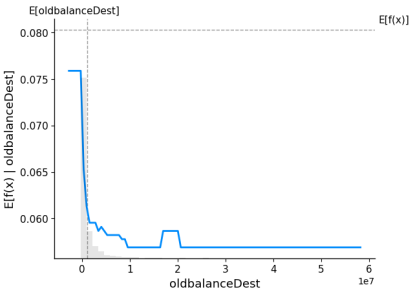

Figure 24. PDP 10%.

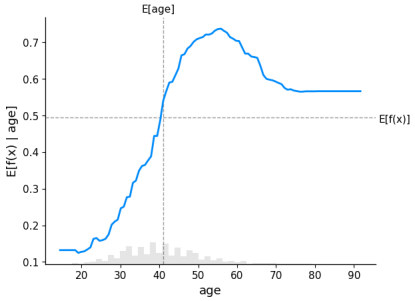

Figure 25. PDP 50%.

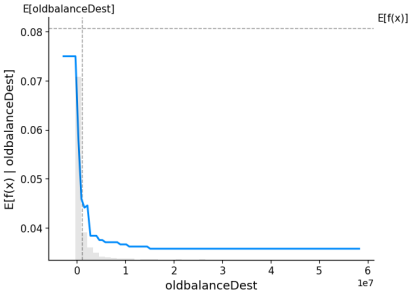

Figure 26. PDP 15%.

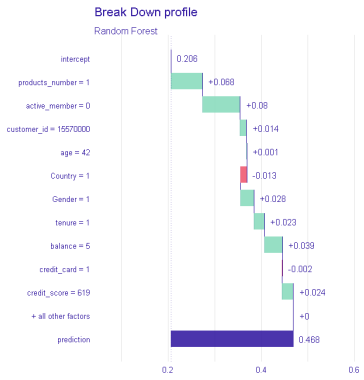

Figure 27. BD 20%.

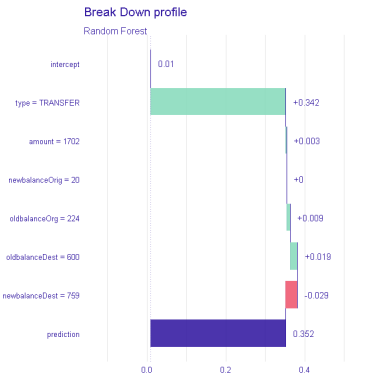

Figure 28. BD 1%.

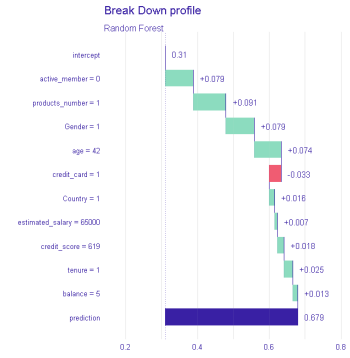

Figure 29. BD 30%.

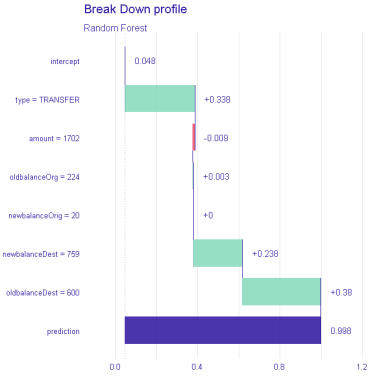

Figure 30. BD 5%.

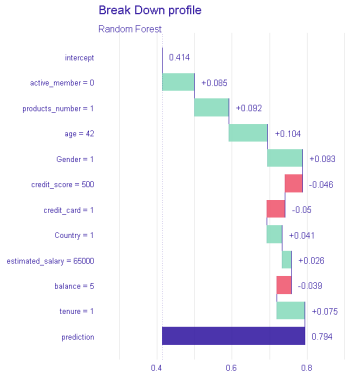

Figure 31. BD 40%.

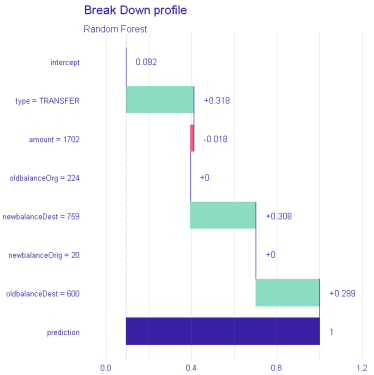

Figure 32. BD 10%.

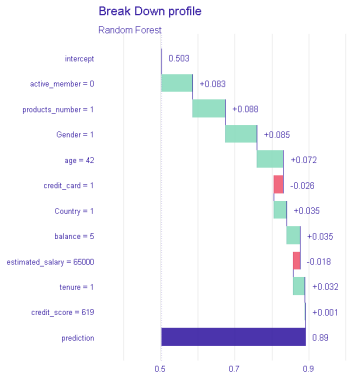

Figure 33. BD 50%.

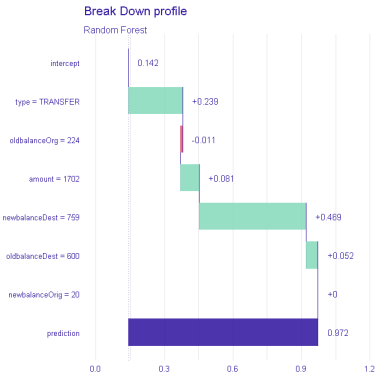

Figure 34. BD 15%.

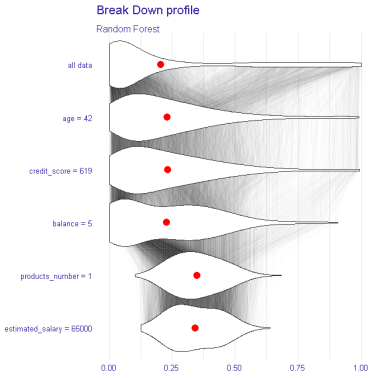

Figure 35. BD 20%.

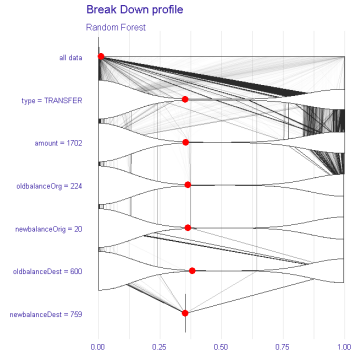

Figure 36. BD 1%.

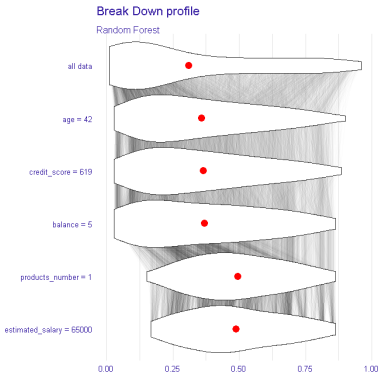

Figure 37. BD 30%.

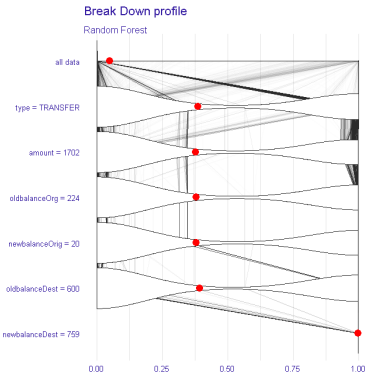

Figure 38. BD 5%.

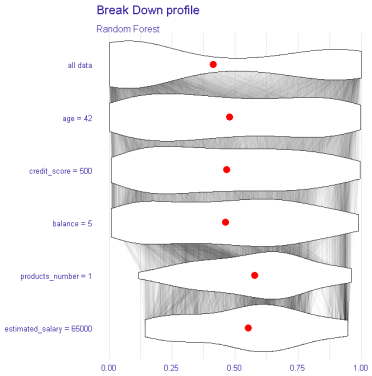

Figure 39. BD 40%.

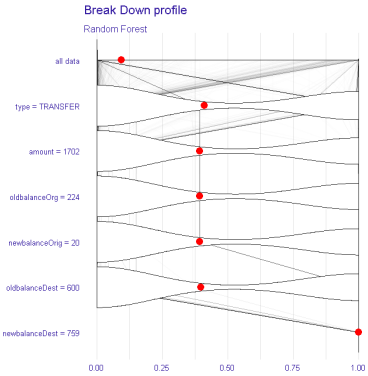

Figure 40. BD 10%.

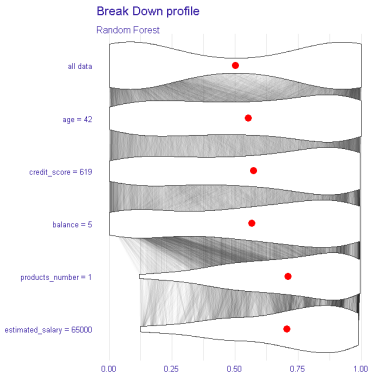

Figure 41. BD 50%.

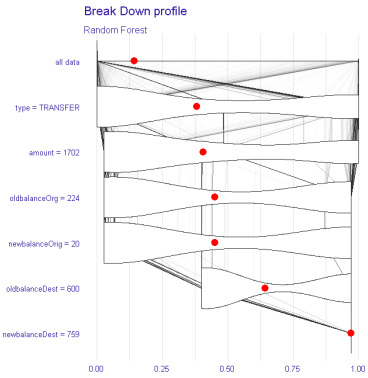

Figure 42. BD 15%.

Next, we looked at the partial dependence plots (PDP) by selected variables in each dataset across various samples of varying class imbalance. In the course of this investigation, the influence of varying class balance on the shape of partial dependence plots (PDPs) was examined using a random forest model. Visual inspection of the PDPs illustrated a consistent overall upward trend in estimated values as both datasets approached a more balanced distribution. A noteworthy observation was the consistent increase in the baseline from 0.12 to 0.5 as the dataset achieved greater balance across the four samples in the churn dataset, according the variable age, see in Figures 19, 21, 23, 25. In the fraud dataset Figures 20, 22, 24, 26, the baseline was as low as below 0.008 at a 1% fraud rate but went as high as 0.08 when the dataset had a 15% fraud rate, according the variable OldbalanceDest. Crucially, the overarching shape of the partial dependence plots remained stable throughout this process. This implies that while the baseline predictions of the model demonstrated an increase with improved class balance, the fundamental relationships between the input variable and the predicted outcome retained their intrinsic characteristics.

We also looked at how individual predictions are affected by class imbalance. Breakdown plots illustrate the manner in which contributions assigned to specific explanatory variables alter the mean model's prediction, resulting in the actual prediction for a particular individual instance or observation. In Figures 27–34, green bars signify positive changes, while red bars represent negative changes in mean predictions, reflecting the contributions attributed to explanatory variables. In Figures 35–42, red dots highlight the mean predictions for the full dataset. Particularly, we were interested in the probability of churn for a male customer aged 42 with a credit score of 619 who earned 65,000 in the churn dataset. To evaluate the impact of imbalance of individual explanatory variables to this particular single-instance prediction, we investigated the changes in the model's predictions when fixing the values of the variables and noted changes as the dataset became more balanced. The two breakdown plots used revealed a significant change in the prediction as the data was more balanced. It can be seen that the predicted value can be as low as below 0.5 when the data is 20% balanced, but can increase the prediction to as high as 0.9 when data is more balanced. This analysis was also followed for the fraud data, and again the predicted probability was as low as 3.5% at a 1% fraud rate and as high as 97% at a 15% fraud rate. In a classification setting, this means that if the model was trained with a wrong class-balance dataset, there is a risk of misclassifying some observations. Similarly, on the average level, this trend was also true for the whole dataset.

8. Discussion

The results of the experiment provided nuanced insights into the intricate relationship between class balance, model performance, and interpretability, particularly in the context of random forest models for churn and fraud detection. One of the most significant findings is the consistent improvement in model performance metrics as class imbalance decreased. This observation aligns with Dube and Verster (2023) and other existing literature on the challenges posed by imbalanced datasets, where the rarity of minority class instances can lead to biased model predictions favoring the majority class. By addressing class imbalance, the experiment demonstrates the potential to mitigate these biases and improve the model's ability to accurately identify rare events such as churn or fraud.

The analysis of Shapley values and feature importance adds depth to our understanding of how individual features contribute to model predictions across varying levels of class imbalance. The

observation that certain features maintain consistent relationships with model predictions regardless of class distribution highlights the importance of these features in capturing meaningful patterns within the data. Conversely, the variability observed in the relationship between other features and model predictions underscores the complexity of feature interactions and their sensitivity to changes in class balance. This insight underscores the importance of considering feature importance in the context of class distribution, as the relevance of features may vary depending on the rarity of the target event. In a similar study done by Chen et al. (2024), it was established that interpretations generated from Shapley values are less stable as the class imbalance increases in a dataset.

Furthermore, the examination of partial dependence plots (PDPs) provided valuable insights into the overall trends in model predictions as class balance improves. Despite variations in baseline predictions, the stability of the underlying relationships between input variables and predicted outcomes suggests robustness in the model's understanding of feature interactions. This finding is particularly significant as it indicates that while class imbalance may influence baseline predictions, it does not necessarily alter the fundamental relationships between features and the target variable. This stability in feature relationships enhances the interpretability of the model and facilitates more informed decision-making.

The analysis of individual predictions through breakdown plots further elucidates the impact of class imbalance on model predictions at the individual level. The observed changes in predicted probabilities highlight the importance of considering class distribution when interpreting individual predictions, as variations in dataset balance can significantly affect the confidence and reliability of model predictions. This insight has practical implications for decision-making in real-world scenarios, where accurate predictions are essential for mitigating risks associated with churn or fraud.

This study represents a pioneering effort in utilizing a comprehensive suite of interpretability tools, including Shapley values, partial dependence plots (PDPs), feature importance analysis, and breakdown plots, to investigate the impact of class imbalance across datasets of varying natures. By integrating these advanced techniques, we bridge a significant gap in the existing literature by offering a holistic understanding of how class imbalance affects model performance and interpretability. This research not only fills a critical void in the current understanding of imbalanced data scenarios but also offers practical insights that can inform the development of more effective and interpretable machine learning models in real-world applications. By closing this gap, our study provides researchers and practitioners with valuable guidance for mitigating the challenges posed by class imbalance and leveraging its potential benefits to enhance predictive accuracy and model interpretability.

In conclusion, the experiment provides valuable insights into the complex interplay between class balance, model performance, and interpretability in random forest models for churn and fraud detection. By elucidating these dynamics, this research contributes to advancing our understanding of effective model development and deployment in scenarios characterized by imbalanced data distributions. These insights have practical implications for improving the reliability and interpretability of machine learning models in real-world applications, particularly in domains where accurate predictions of rare events are critical for decision-making.

9. Conclusions

Our experiment was conducted to explore the impact of class balance on random forest model performance in churn and fraud detection scenarios and has provided valuable insights into the intricate

relationship between data distribution, model performance, and interpretability.

The findings underscore the critical importance of addressing class imbalance in training datasets to enhance the model's ability to accurately identify rare events. The consistent improvement in performance metrics such as precision, recall, F1-score, and accuracy as class imbalance decreases highlights the necessity of balancing the representation of minority and majority classes to achieve optimal predictive performance. Moreover, the analysis of Shapley values and feature importance revealed nuanced insights into the contribution of individual features to model predictions across varying class distributions. While some features exhibited consistent relationships with model predictions, others displayed more variability, emphasizing the complex interplay between feature importance and class distribution. Additionally, the examination of partial dependence plots (PDPs) demonstrated stable trends in estimated values as class balance improved, indicating that fundamental relationships between input variables and predicted outcomes remained unchanged despite variations in baseline predictions. Furthermore, the analysis of individual predictions through breakdown plots emphasized the significant impact of class imbalance on model predictions at the individual level, highlighting the importance of considering class distribution when interpreting model outputs in real-world applications.

Furthermore, while this study provides valuable insights, there are important avenues for future research to explore. Additional methodologies for addressing class imbalance, such as advanced sampling techniques or algorithmic adjustments, warrant investigation to further improve model performance in imbalanced datasets. Moreover, validating the generalizability of these findings across diverse datasets and application domains is essential to ensure the robustness and applicability of the proposed approaches. Additionally, considering the limitations of this study, including the specific characteristics of the datasets used and the choice of machine learning algorithms, future research could benefit from examining alternative models and datasets to provide a more comprehensive understanding of the impact of class imbalance on model performance and interpretability. By addressing these future research directions and considering the study limitations, we can continue to advance the field of imbalanced data analysis and contribute to the development of more effective and reliable predictive models in real-world settings.

Overall, this research contributes to advancing our understanding of the challenges and opportunities associated with imbalanced data in machine learning applications, particularly in domains such as churn and fraud detection. By elucidating the complex interplay between class balance, model performance, and interpretability, this study provides a foundation for developing more robust and reliable predictive models in scenarios characterized by imbalanced data distributions. Moving forward, further research is warranted to explore additional methodologies for addressing class imbalance and to validate the generalizability of these findings across diverse datasets and application domains.

Funding

This work is based on the research supported wholly/in part by the National Research Foundation of South Africa (Grant Number 126885).

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors would like to express their deepest gratitude to their supervisor, Prof. Tanja Verster, for the unwavering guidance, invaluable guidance, and exceptional mentorship throughout the course of this research. They would also like to extend their gratitude to the NWU Centre for BMI for availing their resources to our wonderful staff.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Abd Algani YM, Ritonga M, Bala BK, et al. (2022) Machine learning in health condition check-up: An approach using Breiman's random forest algorithm. *Measurement* 23: 100406.

Ariza-Garzón MJ, Arroyo J, Caparrini A, et al. (2020). Explainability of a machine learning granting scoring model in peer-to-peer lending. *Ieee Access* 8: 64873–64890. https://doi.org/10.1109/ACCESS.2020.2984412

Biecek P, Burzykowski T (2021a) *Explanatory model analysis: explore, explain, and examine predictive models*. CRC Press. https://doi.org/10.1201/9780429027192

Biecek P, Burzykowski T (2021b) Local interpretable model-agnostic explanations (lime). *Explanatory Model Analysis Explore, Explain and Examine Predictive Models*, 1: 107–124.

Breiman L (2001) Random forests. *Mach learn* 45: 5–32. https://doi.org/10.1023/A:1010933404324

Chen Y, Calabrese R, Martin-Barragan B (2024) Interpretable machine learning for imbalanced credit scoring datasets. *Eur J Oper Res* 312: 357–372. https://doi.org/10.1016/j.ejor.2023.06.036

Davis R, Lo AW, Mishra S, et al. (2022) Explainable machine learning models of consumer credit risk. *J Financ Data Sci* 5.

Du Toit H, Schutte WD, Raubenheimer H (2023) Shapley values as an interpretability technique in credit scoring. *J Risk Model Validat* 17.

Dube L, Verster T (2023) Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models. *Data Sci Financ Econ* 3: 354–379. https://doi.org/10.3934/DSFE.2023021

Dube L, Verster T (2024) Assessing the performance of machine learning models for default prediction under missing data and class imbalance: A simulation study. *ORiON* 40: 1–24.

Dumitrache A, Nastu AA, Stancu S (2020) Churn prediction in telecommunication industry: Model interpretability. *J Eastern Eur Res Bus Econ* 2020. https://doi.org/10.5171/2020.241442

Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res* 20: 1–81.

Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. *Pattern Recogn Lett* 27: 294–300.

Goutte C, Gaussier E (2005) A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, 345–359, Springer.

Greenwell BM (2017) pdp: An r package for constructing partial dependence plots. *R J* 9: 421.

Gregorutti B, Michel B, Saint-Pierre P (2017) Correlation and variable importance in random forests. *Stat Comput* 27: 659–678. https://doi.org/10.1007/s11222-016-9646-1

Guliyev H, Tatoğlu FY (2021) Customer churn analysis in banking sector: Evidence from explainable machine learning models. *J Appl Microeconometrics* 1: 85–99.

Hastie T, Tibshirani R, Friedman J, et al. (2009) Random forests. *The elements of statistical learning: Data mining, inference, and prediction*, 587–604.

Jafari MJ, Tarokh MJ, Soleimani P (2023) An interpretable machine learning framework for customer churn prediction: A case study in the telecommunications industry. *J Ind Eng Manage Stud* 10: 141–157. https://doi.org/10.22116/jiems.2023.365114.1504

Jiao Y, Du P (2016) Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant Biol* 4: 320–330. https://doi.org/10.1007/s40484-016-0081-2

Liaw A, Wiener M, et al. (2002) Classification and regression by randomforest. *R News* 2: 18–22.

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Adv Neur Inf Process Syst* 30.

Moraffah R, Karami M, Guo R, et al. (2020) Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explor Newsl* 22: 18–33. https://doi.org/10.1145/3400051.3400058

Nationalbank Oesterreichische (2004). *Guidelines on credit risk management: Rating models and validation*. Oesterreichische Nationalbank.

Nohara Y, Matsumoto K, Soejima H, et al. (2022) Explanation of machine learning models using Shapley additive explanation and application for real data in hospital. *Comput Meth Prog Bio* 214: 106584.

Peng K, Peng Y, Li W (2023) Research on customer churn prediction and model interpretability analysis. *Plos one* 18: e0289724.

Ribeiro MT, Singh S, Guestrin C (2016) Model-agnostic interpretability of machine learning. *arXiv preprint*. https://doi.org/10.48550/arXiv.1606.05386

Rodríguez-Pérez R, Bajorath J (2019) Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. *J Med Chem* 63: 8761–8777.

Shahhosseini M, Hu G (2021) Improved weighted random forest for classification problems. In *Progress in Intelligent Decision Science: Proceeding of IDS 2020*, 42–56, Springer.

Shapley L (2020) A value for n-person games. *Class Game Theory* 69–79.

Staniak M, Biecek P (2018) Explanations of model predictions with live and breakdown packages. *arXiv preprint*.

Tekouabou SC, Gherghina SC, Toulni H, et al. (2022) Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods. *Mathematics* 10: 2379. https://doi.org/10.3390/math10142379

Tran KL, Le HA, Nguyen TH, et al. (2022) Explainable machine learning for financial distress prediction: evidence from Vietnam. *Data* 7: 160. https://doi.org/10.3390/data7110160

Uddin MS, Chi G, Al Janabi MA, et al. (2022) Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability. *Int J Financ Econ* 27: 3713–3729. https://doi.org/10.1002/ijfe.2346

Verster T, Fourie E (2023) The changing landscape of financial credit risk models. *Int J Financ Stud* 11: 98. https://doi.org/10.3390/ijfs11030098

Winham SJ, Freimuth RR, Biernacka JM (2013) A weighted random forests approach to improve predictive performance. *Stat Anal Data Min ASA Data Sci J* 6: 496–505. https://doi.org/10.1002/sam.11196

Yu F, Wei C, Deng P, et al. (2021) Deep exploration of random forest model boosts the interpretability of machine learning studies of complicated immune responses and lung burden of nanoparticles. *Sci Adv* 7: eabf4130. https://doi.org/10.1126/sciadv.abf413

Zhu X, Chu Q, Song X, et al. (2023) Explainable prediction of loan default based on machine learning models. *Data Sci Manag* 6: 123–133. https://doi.org/10.1016/j.dsm.2023.04.003

AIMS Press