Ensemble Counterfactual Explanations for Churn Analysis

Samuele Tonati[1,2(✉)], Marzio Di Vece[1,3], Roberto Pellungrini[1], and Fosca Giannotti[1]

[1] Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy
{samuele.tonati,marzio.divece,roberto.pellungrini,fosca.giannotti}@sns.it
[2] University of Pisa, Lungarno Antonio Pacinotti 43, 56126 Pisa, Italy
[3] IMT School for Advanced Studies, Piazza San Francesco 19, 55100 Lucca, Italy

Abstract. Counterfactual explanations play a crucial role in interpreting and understanding the decision-making process of complex machine learning models, offering insights into why a particular prediction was made and how it could be altered. However, individual counterfactual explanations generated by different methods may vary significantly in terms of their quality, diversity, and coherence to the black-box prediction. This is especially important in financial applications such as churn analysis, where customer retention officers could explore different approaches and solutions with the clients to prevent churning. The officer's capability to modify and explore different explanations is pivotal to his ability to provide feasible solutions. To address this challenge, we propose an evaluation framework through the implementation of an ensemble approach that combines state-of-the-art counterfactual generation methods and a linear combination score of desired properties to select the most appropriate explanation. We conduct our experiments on three publicly available churn datasets in different domains. Our experimental results demonstrate that the ensemble of counterfactual explanations provides more diverse and comprehensive insights into model behavior compared to individual methods alone that suffer from specific weaknesses. By aggregating, evaluating, and selecting multiple explanations, our approach enhances the diversity of the explanation, highlights common patterns, and mitigates the limitations of any single method, offering to the user the ability to tweak the explanation properties to their needs.

Keywords: Explainable AI · Counterfactual Explanations · Churn Analysis

1 Introduction

The recent surge of interest in Machine Learning (ML) and Artificial Intelligence (AI) has led to the development of a multitude of models aimed at decision-making across various sectors, including healthcare, financial systems, and criminal justice. Although it may seem logical to favor more accurate models when

D. Pedreschi et al. (Eds.): DS 2024, LNAI 15244, pp. 332–347, 2025.
https://doi.org/10.1007/978-3-031-78980-9_21

evaluating different options, the emphasis on accuracy has resulted in unintended consequences. ML Developers frequently prioritize higher accuracy, often at the expense of interpretability, making their models increasingly complex and difficult to understand. This lack of explainability becomes a significant concern when models are entrusted with making critical decisions that affect people's well-being. In order to address these concerns, the concept of Explainable AI (XAI) has emerged as a promising solution. XAI addresses the aforementioned challenges by offering explanations to make the inner workings of AI models interpretable and easy to understand, as AI-generated prescriptions are often unintelligible to humans. Counterfactual Explanations are a ubiquitous form of explanation, which aim to provide a contrastive way of explaining the decisions of a model. They also align with the requirements specified by the GDPR and the AI Act, in particular concerning the right to explanation and transparency. Counterfactual Explanations indicate what would need to change in the input data to alter the model's output, thereby offering a clear and intuitive understanding of the decision-making process and are found to be successful in many practical domains [1,2]. This is especially true for applications such as churn analysis. In such a scenario we have a churn officer relying on a machine learning model to understand which clients are more likely to churn. Here, understanding why certain clients are churning and being able to provide feasible alternatives to prevent churn is the key objective [3]. Therefore, explainable AI techniques are fundamental in giving some insight into what actions the officer could employ to engage the customer and prevent churn [4]. While there are many techniques to generate counterfactual explanations, each methodology has its own specific mechanisms that may provide counterfactuals with specific characteristics. Moreover, there is no accepted methodology in the literature to evaluate counterfactual explanations agnostically w.r.t. the application domain, therefore, the counterfactual explanations obtained with a certain methodology may have properties that are not entirely helpful in the context of churn analysis. In this paper, we propose a novel evaluation framework for counterfactual explanations for churn analysis. We rely on an ensemble of counterfactual explanations generated with diverse techniques and a flexible linear combination of metrics as an evaluation function to select the best counterfactuals. This function allows the officer to modify the counterfactual ensemble with little computational overhead, selecting the counterfactuals that could give the best chances of retention. We evaluate our approach on three publicly available datasets for churn prediction.

2 Related Works

Churn analysis is a crucial application of ML in various industries, particularly in telecommunications, finance, and subscription-based services. Churn refers to the phenomenon where customers stop using a company's services or products, leading to a loss of revenue [5–7]. Accurately predicting churn is essential for businesses to implement effective retention strategies and reduce customer attrition. Traditional ML models used for churn prediction focus on identifying patterns

and factors that contribute to a customer's likelihood of leaving and are designed often as a combination of unsupervised and supervised techniques [8–11]. These models often rely on large datasets encompassing customer behavior, transaction history, demographics, and other relevant attributes. While such models can achieve high accuracy in predicting churn, their complexity and opaqueness in their decision-making processes poses challenges for decision-makers who need to understand the underlying reasons for customer attrition (i.e. the loss of customers by a business).

Counterfactual explanations are a distinct category of post-hoc local explanation methods that describe how to alter the input to achieve a desired outcome from the model. According to a commonly accepted definition in literature, these modifications should be minimal and closely resemble the original instance being explained. This similarity highlights how counterfactual explanations share many traits with adversarial attacks, as both aim to flip the model's prediction, albeit with different goals. The key distinguishing factor of counterfactual explanations lies in their desired properties, which are intended to provide informative value to the user. These properties typically include proximity to the original instance, minimality, actionability, and diversity among the others [2].

Counterfactual explanations align well with the goal of personalized customer experiences [12,13]. By understanding the specific factors that influence each customer's decision to stay or leave, businesses can tailor their engagement strategies to meet individual needs, thus enhancing customer satisfaction and loyalty. Moreover, counterfactual explanations have yet to be applied to the case of churn in the literature. Existing works have instead mainly focused on XAI techniques based on feature importance [4,14,15].

3 Problem Definition

Churn prevention encompasses all the actions that a company puts into place to prevent the loss of customers. The first and most important part of any churn prevention strategy lies in detecting which customers will likely interrupt their relationship with the company, given their current status. This task can be modeled as a binary classification task [3] where a machine learning model is used to predict which customers are the ones likely to churn. A churn officer has then the duty of interacting with these customers in order to find possible actions to prevent them from churning.

In this context, counterfactual explanations appear to be extremely useful from a business' perspective: firstly, counterfactual explanations help identify the minimal changes needed to retain a customer; for example, if the model indicates that a customer is likely to churn, a counterfactual explanation might reveal that offering a small discount, offering a particular product or improving service quality could change the prediction to retention. This allows businesses to implement precise interventions that are cost-effective and efficient. Secondly, the interpretability offered by counterfactual explanations builds trust among business stakeholders. Unlike explanatory methods that might provide abstract

or general insights, such as in the case of global explanations, counterfactuals show specific scenarios and outcomes, making it easier for non-technical stakeholders to understand and trust the model's recommendations. This trust is vital in securing support and commitment from stakeholders for data-driven strategies and ensuring their successful implementation. However, depending on what counterfactual method one chooses, the explanations obtained may rely on specific optimization strategies and therefore explore only some particular aspect of the importance of a churn officer. Counterfactual explanations have not been explored as a solution to this kind of problem [4,14,15]. To tackle similar problems, Guidotti et al. [16] proposed an ensemble method that leverages the strengths of multiple counterfactual explainers to cover a set of desirable properties, such as minimality, actionability, stability, diversity, plausibility, and discriminative power. Their approach demonstrates the efficacy of boosting weak explainers into a powerful ensemble that is both model-agnostic and data-agnostic, capable of handling various data types including tabular data, images, and time series. Building upon this idea, we propose an ensemble approach that operates ex-post as an evaluation and selection mechanism. Our method is designed to identify the optimal set of counterfactual examples by employing a linear combination score of various metrics, that reflect on the possible aspects that a churn officer would explore in a churn prediction model. In contrast to the ensemble proposed by [16], which combines results through a diversity-driven selection function, our framework introduces a more nuanced evaluation score. This approach not only refines the selection process but also ensures that the chosen counterfactuals align closely with the desired properties - thereby improving the interpretability and reliability of the explanations provided - and that can be aptly tweaked by practitioners to give more emphasis to a specific metric. In the context of a binary classification task, given a model function $f : X \subset \mathbb{R}^n \rightarrow \{0, 1\}$ that maps instances $x \in X$ with features in $\mathbb{R}^n$ to predicted class labels (0 or 1), a counterfactual explanation aims to identify a modified version of an original instance x such that the modified instance x' leads to a different model prediction. Therefore, given an original instance ($x \in \mathbb{R}^n$), a predictive model ($f : \mathbb{R}^n \rightarrow \mathbb{R}$), the model's prediction for the original instance ($y_M = f(x)$) and a target prediction ($y_T \neq y_M$), The goal is to find a counterfactual instance ($x' \in \mathbb{R}^n$) such that $f(x') = y_T$ while ensuring that (x') is similar to the original instance (x). This similarity can be a function of the different metrics evaluating the relationships between these two instances.

Let's consider the case in which the similarity function is represented by a distance metric $d(x, x')$, typically chosen as the (L_1) or (L_2) norm. The problem of finding a counterfactual explanation can be solved by finding $\min_{x'} d(x', x)$ subject to the constrain $f(x') = y_T$ where $d(x', x)$ is defined as $d(x', x) = \|x' - x\|_p$ with ($\|\cdot\|_p$) denoting the (L_p)-norm, commonly (L_1) or (L_2)-norm.

Other than distance metrics, the optimization process may also incorporate constraints on the perturbation ($\delta_{CF} = x' - x$) to ensure sparsity or adherence to feature-specific constraints: $x' = x + \delta_{CF}$.

Our proposal is to score the counterfactual explanations produced by an ensemble of counterfactual methods using evaluation metrics that align with desired properties in the context of Churn analysis. The pseudocode of our approach is given in Algorithm 1.

Algorithm 1. k-CEM: k-Counterfactual Ensemble Method

1: **Input:** X_{test} - test set, E - set of CF explanations, M - trained model
2: **Output:** C - ensemble of top counterfactual explanations
3: $C \leftarrow \emptyset$ ▷ Initialize result set
4: **procedure** EnsembleCFExplanations
5: Load X_{test}, E, and M
6: **for** each $e \in E$ **do**
7: $C \leftarrow$ Extract subset of X_{test} corresponding to e
8: $\hat{y}_{orig} \leftarrow M(C_{subset})$
9: $\hat{y}_{cf} \leftarrow M(e)$
10: **end for**
11: Retain CFs where $\hat{y}_{test} \neq \hat{y}_{ensemble}$
12: **for** each CF $c \in C$ **do**
13: Calculate and normalize metrics: $d_{prox}(c)$, $s(c)$, $p(c)$, $d_{div}(c)$
14: Define weights $w_{prox}, w_{spars}, w_{plaus}, w_{div}$
15: Compute score: $score(c) \leftarrow w_{prox} \cdot d_{prox}(c) + w_{spars} \cdot s(c) + w_{plaus} \cdot p(c) + w_{div} \cdot (1 - d_{div}(c))$
16: **end for**
17: Sort C by *score* in ascending order
18: Group C by original instance index i
19: **for** each group g_i **do**
20: Select top k CFs from g_i based on *score*
21: **end for**
22: Return the top CFs DataFrame C
23: **end procedure**

As described in the pseudocode our method takes in input test data, counterfactual explanations, and a trained model. For each explanation method we retain only valid explanations and put them in an ensemble set C (lines 6–11). Then, we characterize each element of the set using relevant metrics which are first normalized and then incorporated into the score function with user-defined weights (lines 12–16). Explanations are then sorted by their score and the top k are selected for each reference instance in the test set (lines 17–22).

3.1 Counterfactuals Methods

For the implementation of our Counterfactual Ensemble Method, we choose four different counterfactual generation methods that, in our opinion, condense the most diverse approaches to synthetic counterfactual explanation generation.

- **DiCE** perturbs input features within model decision boundaries, leveraging a genetic algorithm to create multiple instances leading to different predictions [17]. It generates diverse counterfactual examples solving an optimization problem that balances properties of proximity and diversity.
- **Growing Spheres** (GS) uses a sphere-growing algorithm to iteratively explore the feature space around a given instance [18]. In our approach, we slightly modify GS to return the best k instances instead of just one, ranking them based on L_2 proximity to the original instance.
- **CFRL** is a model-agnostic counterfactual generation method that uses reinforcement learning [19] to train a generative model to produce counterfactual explanations.
- **T-LACE** is a counterfactual explanation method that constructs a transparent latent space using a linear transformation where also the original prediction of the model is added, ensuring that similar records in the latent space have similar features and predictions [20]. Counterfactuals are then searched in the latent space decomposing contributions from each feature to identify a prediction direction.

3.2 Evaluation of Counterfactual Explanations

The scoring function of the counterfactual ensemble is the core of our approach, and it is based on properties that reflect upon possible questions that a churn officer would need to answer in order to prevent a client from churning. Here, we present the measures we chose and highlight their purpose in the context of churn analysis.

Proximity Measures. *How minimal are the changes required to retain potentially churning customers? Proximity measures indicate close counterfactuals.* We choose an average proximity measure using a geometric mean that combines various normalized proximity measures. The geometric mean prevents skewing by outliers, ensuring equal contribution from all proximity measures. The individual proximity metrics we use are:

Euclidean Distance (L_2 norm) measures the overall difference in feature values:

$$\text{Proximity}_{L2} = \sqrt{\frac{m-h}{m} \sum_{i \in \text{cont}} (x'_i - x_i)^2 + \frac{h}{m} \sum_{j \in \text{cat}} \delta(x'_j, x_j)}$$

Manhattan Distance (L_1 norm) measures the sum of absolute differences:

$$\text{Proximity}_{L1} = \frac{m-h}{m} \sum_{i \in \text{cont}} |x'_i - x_i| + \frac{h}{m} \sum_{j \in \text{cat}} \delta(x'_j, x_j)$$

Maximum Absolute Difference L_∞ norm measures the maximum element-wise absolute difference:

$$\text{Proximity}_{L\infty} = \max\left(\frac{m-h}{m}\max_{i\in\text{cont}}|x'_i - x_i|, \frac{h}{m}\max_{j\in\text{cat}}\delta(x'_j, x_j)\right)$$

Here, m is the total number of features, h the number of categorical features, cont continuous features, cat categorical features, and $\delta(x'_j, x_j)$ is 1 if $x'_j \neq x_j$ and 0 otherwise (Hamming distance).

Plausibility Measure. *Is the counterfactual explanation similar to a non-churning customer in the data and thus justifiable to the customer? Plausibility indicates counterfactuals that have close examples in the original dataset.* The plausibility measure assesses the degree of plausibility or soundness of the counterfactual instances (X') with instances in the original dataset to explain.

Specifically, it calculates the minimum distance of each $x' \in X'$ from its closest instance in the original data.

To compute the plausibility measure we build a KDTree on the X_{test} dataset to efficiently find the nearest neighbors, then we query the KDTree to find the nearest neighbor in X_{test} for each instance in the set of counterfactual instances and we calculate the distance between each x' and its closest instance in X_{test}. The use of a KDTree for computing the plausibility measure is motivated by the need to efficiently find the nearest neighbors of counterfactual instances within the dataset X_{test}. KDTree provides logarithmic search time complexity for nearest neighbor queries, making it more scalable compared to linear search methods, which have linear time complexity. Efficiency is crucial when the number of instances in X_{test} is large, a common situation for real-world applications like churn analysis.

Plausibility is represented as the euclidean distance of x' from its closest instance in the X_{test} population.

Sparsity Measure. *Does the counterfactual explanation modify as few features as possible, thus making the required changes easier for the churn officer to propose? Sparsity indicates counterfactuals that touch the least amount of features.* Sparsity is computed as the fraction of differing features to the total number of features n:

$$\text{Sparsity} = \frac{\sum_{i=1}^{n}(x'_i \neq x_i)}{n} \tag{1}$$

Diversity Measure. *The counterfactuals produced do provide different courses of action for the churn officer? Diversity indicates that the produced explanations have enough variety for the churn officer to act upon.* The diversity measure quantifies the dissimilarity or variation within groups defined by the generation source. It is calculated as the mean of distances between pairs of instances within each group.

$$\text{Diversity} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i(n_i-1)}\sum_{j\neq k} d(x_j, x_k) \tag{2}$$

where N is the total number of groups - i.e. sets of counterfactuals for a given instance to explain -, n_i is the number of instances in group i, and $d(x_j, x_k)$ is the distance between instance j and instance k within the same group.

Evaluation Score. Our evaluation score is computed by combining multiple measures using a weighted linear combination. This approach aims to synthesize different aspects of the counterfactual explanations into a single score, allowing for comprehensive evaluation and comparison. The first three metrics need to be minimized-i.e., the lower, the better, whereas, since diversity needs to be maximized, we minimize its complement.

$$\begin{aligned}\text{Evaluation Score} = &+ w_1 \times \text{Proximity}_{\text{avg;N}} \\ &+ w_2 \times \text{Sparsity} \\ &+ w_3 \times \text{Plausibility}_{\text{N}} \\ &+ w_4 \times (1 - \text{Diversity}_{\text{N}})\end{aligned} \tag{3}$$

Each weight represents the relative importance assigned to its corresponding measure in the overall evaluation. The selection of top counterfactual examples involves sorting the instances based on their linear combination scores in ascending order. Instances with lower scores are prioritized as they represent better adherence to the optimization objectives defined by the weighted combination of measures. By leveraging this approach, churn experts can efficiently identify and retrieve the most relevant and effective counterfactual explanations for each individual instance, in any scenario.

4 Experiments

4.1 Introduction

For our experiments, we used three public datasets specifically focused on the churn classification problem. The "Churn for Bank Customers" dataset[1] contains 10,000 records with 14 features aimed at predicting whether a customer has exited the bank (0.20 ratio of churners). The "Credit Card Bank Churn", dataset[2] includes 10,000 credit card user records with 18 features to predict if a customer will stop using the bank's credit card services (0.19 ratio of churners). The "Iranian Churn Dataset"[3] contains 3,150 records and provides telecommunications customer data from Iran with 13 features used to analyze churn behavior (0.18 ratio of churners) in the telecom industry.

[1] https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers.
[2] https://www.kaggle.com/datasets/anwarsan/credit-card-bank-churn.
[3] https://archive.ics.uci.edu/dataset/563/iranian+churn+dataset.

We start by identifying which model to explain. We compare the performance of Light Gradient Boosting Machine (LightGBM)[21], XGBoost [22], and Random Forest classifiers [23]. Our focus is on the explanation methodology, more so than on the task, we therefore chose models that are commonly used for the task and easily applied [3]. To fine-tune the models, we conducted a randomized grid search with 5-fold cross-validation, optimizing for the ROC AUC score and we accounted for class imbalance penalizing errors on the minority class proportionally during training.

In Table 1 the comparison is displayed across datasets for F1-Score and Matthews Correlation Coefficient (MCC) measures. The LightGBM outperforms or at least performs as good as the XGBoost in the datasets under analysis while the Random Forest slightly underperforms. These results lead us to establish the LightGBM as the model of interest for the following counterfactual explanations. Statistics related to counterfactual explanations for the XGBoost model are similar and will be omitted to enhance clarity.

Table 1. Model Performance on Churn Datasets. LightGBM and XGBoost display similar performances while Random Forest slightly underperforms.

Metric	Bank Churn			Card Churn			Iranian Churn		
	2000 instances			2026 instances			630 instances		
	LGB	XGB	RF	LGB	XGB	RF	LGB	XGB	RF
F1 Score	0.60	0.59	0.57	0.87	0.86	0.80	0.90	0.88	0.88
MCC	0.52	0.48	0.50	0.79	0.78	0.73	0.87	0.74	0.85

4.2 Exploring Weight Combinations for Counterfactual Evaluation

We explore different weight combinations of our scoring function for evaluating counterfactual explanations. Specifically, we take into consideration the possibility of assigning equal weights on the four metrics (0.25 for each metric), or the possibility of focusing on one of them by imposing a "Higher Weight" (HW) on it, i.e. a weight of 0.5, while imposing for the others a weight of 0.1667.

For each weight combination, we calculate the score for each counterfactual example and select the top $k = 5$ examples for each algorithm. The counts of top counterfactuals selected in the ensemble method for each algorithm against the weight combinations are plotted in the figure below.

We can see in Fig. 1 that for the Bank Churn and the Iranian Churn datasets, the counterfactuals produced using the DiCE method are predominantly selected, regardless of the imposed imbalance on the score function. Contrarily, the T-LACE method is predominantly selected in the generation of counterfactuals for the Credit Card Churn dataset. The quality of counterfactuals across methods is, hence, strongly dataset-dependent. This shortcoming can be avoided by the proposed ensemble approach, being capable of selecting

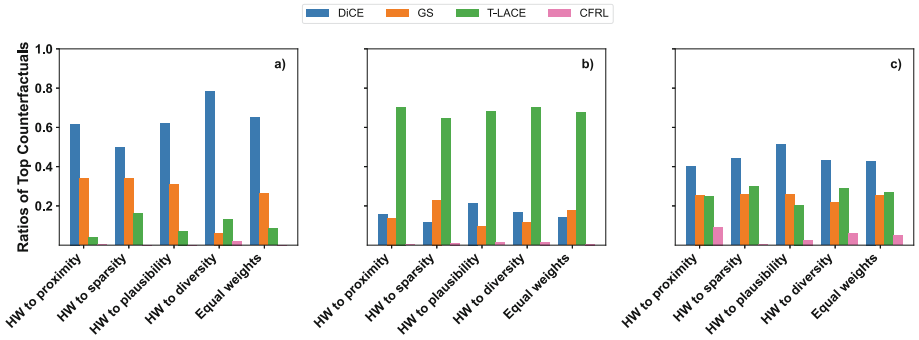


Fig. 1. Ratio of top counterfactuals by weight combinations that are selected by the ensemble for each of the three datasets: **a)** the Churn for Bank Customers, **b)** the Credit Card Bank Churn and **c)** the Iranian Churn dataset. In the x-labels, *HW* stands for "higher weight" in the score function (default value 0.5). While the ensemble selects a relatively large amount of counterfactuals originating from the DiCE method for **a)** and **c)**, it selects predominantly counterfactuals from T-LACE for Credit Card Churn. The ensemble method, hence, displays a high degree of adaptability to the dataset at hand.

at the individual-instance level, the best-performing counterfactual explanations according to the score function imposed by the user (following her preferences).

Next, we explore different weight combinations and predicted probability thresholds for evaluating counterfactual explanations. As for weight combinations, we refer to the ones introduced in Fig. 1, as for predicted probability threshold, instead, we refer to 3 different classes of thresholds for the predicted probability that the counterfactual instance has the opposite class of the reference instance according to the black-box model. The thresholds considered are 0.5, 0.7, and 0.9. For each combination of weights and thresholds, we calculate the scoring function for each counterfactual example and select the top 5 counterfactuals based on their scoring. The counts of top counterfactuals for each algorithm are then plotted against the weight combinations and predicted probability thresholds, as shown in Fig. 2.

As we can see from Fig. 2, in the Bank Churn (a) and Iranian Churn datasets (c), GS and CFRL algorithms tend to be selected, together with the predominant DiCE, for counterfactuals with lower confidence, i.e. prediction probability $\in (0.5, 0.7)$. Increasing the prediction confidence of the top counterfactuals offered by GS and CFRL reduce in favor of counterfactuals generated by DiCE. Instead, in the Credit Card Churn dataset, T-LACE is predominant across probability thresholds, while GS and DiCE offer a number of counterfactuals that fluctuate regardless of the given threshold. This shows how it is possible to explore the confidence of the underlying model by simply modifying the scoring function.

The plots in Fig. 3 offer insights into the performance of counterfactual explanations generated from different sources. We observe a peculiar trait in our ensemble: prediction probabilities are more spread-out, with respect to the sin-

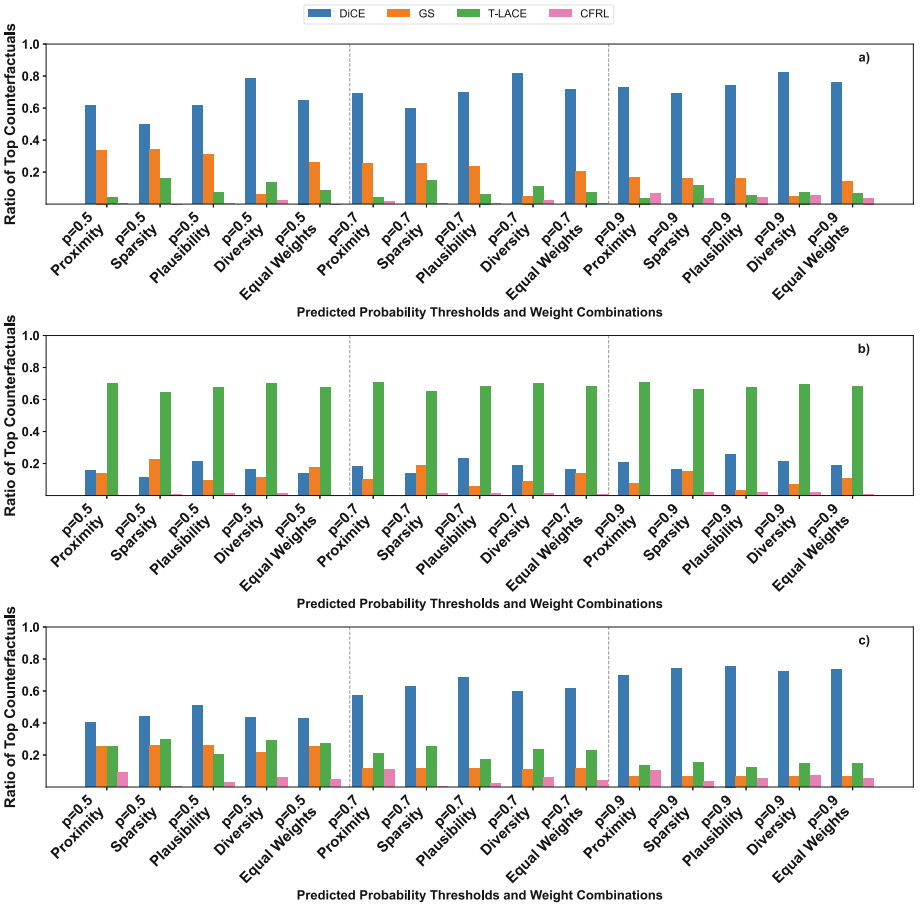
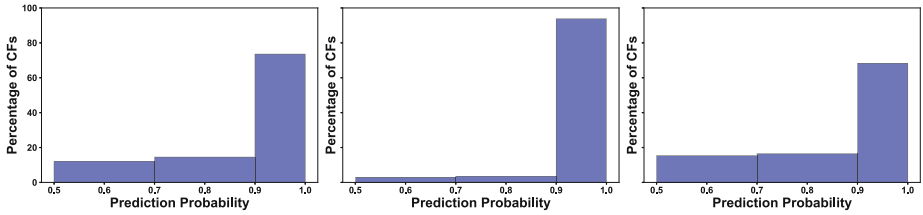
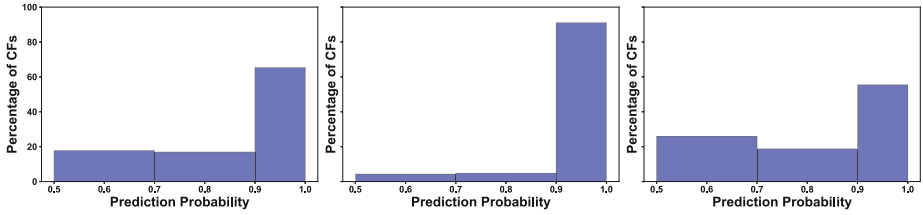


Fig. 2. Ratio of top counterfactuals by probability threshold and weight combinations for each of the three datasets: **a)** Churn for Bank Customers, **b)** Credit Card Bank Churn and **c)** Iranian Churn. In the x-labels, different thresholds are depicted stratified by the imposed coefficient on the score function. In the Credit Card Churn dataset, T-LACE is the predominant method across thresholds. In the other datasets, the predominant method is DiCE. However, note that in both datasets the distance between the ratio of top counterfactuals between DiCE and the other methods increases by increasing the threshold, in other words the selection of counterfactuals across the different methods tends to be more consistent when prediction confidence is lower.

gle methods. This is intriguing because it provides a set of explanations that can cover all the ranges of churning probability (i.e. <0.7: low churn probability, <0.9: mid churn probability, and >0.9: high churn probability), offering a possibility for market segmentation and diverse intervention strategies. A customer retention specialist might get a more exhaustive explanation with diversity in prediction probabilities rather than a focus only on labels with high confidence.



(a) Aggregated single methods



(b) Ensemble method

Fig. 3. Comparison of prediction probability of the single methods aggregated together (a) with the ensemble method (b) for the different datasets, namely Bank Churn (on the left), Credit Card Churn (on the center), and Iranian Churn (on the right). The prediction probability given by the application of the black-box on the counterfactual instances and the percentage of CFs with those probabilities is displayed respectively on the x-axis and the y-axis. Plots of the ensemble in (b) show higher diversity in probabilities with respect to single methods in (a). The increased diversity in prediction probabilities allows practitioners to better segment counterfactuals, and hence individuals, based on different confidence levels.

4.3 Analysis of Features Change Ratios

The plot in Fig. 4 displays a heatmap of Kendall's Tau correlations between the SHAP features rank and the ranks of the most changed features from the ensemble method, listed in decreasing order. SHAP values measure the contribution of each feature to the prediction for an individual instance, based on the concept of Shapley values [24]. These values help to understand the influence of each feature on the model's output.

The average change ratios of the features, on the other hand, indicate the proportion of counterfactual instances where a feature value differs from the original instance. This metric helps to identify which features are most frequently altered in the generated counterfactuals, suggesting their importance in driving changes in predictions. This visualization provides insights into how well the ranks of the SHAP values align with the ranks of the most changed features.

As illustrated in Fig. 4, methods like T-LACE and DiCE exhibit inconsistent correlations across different datasets which suggests they might be sensitive to specific dataset characteristics. On the other hand, GS is more consistent but uncorrelated and CFRL shows consistently negative correlations. The ensemble in all its variations frequently shows stronger and more consistent correlations,

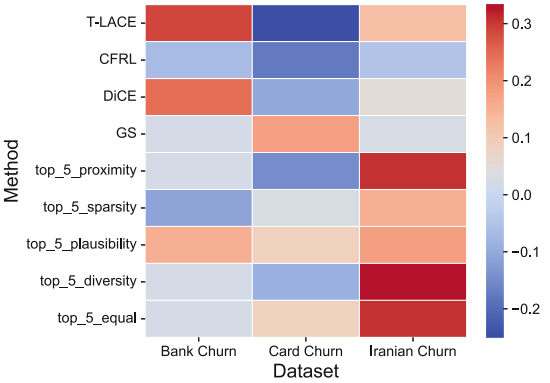


Fig. 4. Kendall-Tau heatmap for SHAP rank correlations, comparing the single methods to the ensemble with different weights configurations across the datasets. Results show that the ensemble selects diverse and more consistent CFs that have on average higher correlation to SHAP features rank.

indicating its robustness in aligning counterfactual changes with SHAP feature importance. The varying correlations suggest that different methods for generating counterfactual explanations may prioritize different features compared to SHAP. Methods with higher positive correlations - as in the case of the ensemble - can be considered more interpretable and aligned with SHAP's feature importance, making them preferable in scenarios where feature importance is a key part for effective actionability of the explanations.

4.4 Example of the Explanation

In Table 2 the output of the ensemble method is displayed and compared with a chosen original instance to be explained. Only the changes to the features made by the generation methods to flip the decision boundary of the black-box are reported. In this example, the original observation is a churning customer, and the ensemble offers a diverse set of counterfactual instances to explain alternative scenarios where the customer would not be predicted as churning. The user can explore different options, a more sparse alternative as in the case of CF_1, a more plausible (i.e. more similar to an actual example in the data) option as in CF_2, a low effort option as in CF_3 or a more diverse one as in CF_5.

Table 2. Comparison of the original instance with the set of top 5 counterfactual instances. In the chosen example from Bank Churn Dataset, CF_1 and CF_3 are selected from GS, CF_2 and CF_4 from DiCE and CF_5 from T-LACE.

Original Instance

CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Churn
650	1	30	6	0	1	0	0	67997	1

Counterfactual Instances

CF_1:	Tenure +1 → Churn = 0
CF_2:	CreditScore -157, NumOfProducts +1, EstimatedSalary -2181 → Churn = 0
CF_3:	HasCrCard +1 → Churn = 0
CF_4:	CreditScore -300, NumOfProducts +1, EstimatedSalary -2807 → Churn = 0
CF_5:	Tenure +3, EstimatedSalary -67997 → Churn = 0

5 Conclusions

In this paper we focused on a counterfactual ensemble strategy for churn analysis, leveraging an evaluation score specifically suited to enhance the counterfactual explanation's utility for churn officers and experts. Our method is applicable to any type of classifier, regardless of its specific characteristics, as it only requires the classifier's prediction during the counterfactual generation phase. The ensemble method leverages the strengths of multiple counterfactual generation techniques, and the evaluation function ensures that useful properties are prioritized for the churn officer. This results in a more diverse set of explanations that offer a broader perspective on the local decision boundary of the black-box and multiple courses of intervention. The proposed evaluation scoring function allows for a multifaceted evaluation of counterfactual explanations based on multiple criteria which we believe is particularly useful in the context of churn analysis, where counterfactual explanations can provide actionable insights that can help businesses devise targeted intervention strategies. For example, understanding the minimal changes needed to retain a customer can lead to cost-effective and efficient retention strategies. The ability to tweak the evaluation parameters based on specific needs allows practitioners to prioritize certain aspects of the explanations, such as minimizing changes or ensuring high plausibility, making the approach highly adaptable to various scenarios. The experiments conducted on three publicly available churn datasets from different domains (banking, credit card services, and telecommunications) have shown the applicability of our evaluation strategy, and how easily it can be piloted by churn officers to explore churn prediction machine learning models. While the current implementation of our work has shown promising results, there are several areas for future research and improvement. Conducting user studies to evaluate the business utility and user satisfaction of the generated counterfactual explanations would provide valuable insights for refining the approach. Moreover, the computational costs associated with counterfactuals generation must be carefully considered when dealing with very large datasets, as it may hinder efficiency in real-time churn analysis scenarios. Additionally, exploring the integration with other explainability techniques,

such as global explanations and feature importance measures, could provide a more comprehensive toolkit for understanding model behavior.

Acknowledgements. SoBigData.it receives funding from European Union - NextGenerationEU - National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) - Project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013 - Avviso n. 3264 del 28/12/2021. This work has been also supported by the PNRR-M4C2-Investimento 1.3, Partenariato Esteso PE00000013-"FAIR-Future Artificial Intelligence Research"-Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme. MDV also acknowledges support by the European Community programme under the funding schemes: ERC-2018-ADG G.A. 834756 "XAI: Science and technology for the eXplanation of AI decision making." This work was also funded by the European Union under Grant Agreement no. 101120763 - TANGO. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

References

1. Stepin, I., Alonso, J.M., Catalá, A., Pereira-Fariña, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access **9**, 11974–12001 (2021)
2. Guidotti, R.: Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery (2022)
3. Geiler, L., Affeldt, S., Nadif, M.: A survey on machine learning methods for churn prediction. Int. J. Data Sci. Anal. **14**(3), 217–242 (2022)
4. Joy, U.G., Hoque, K.E., Uddin, M.N., Chowdhury, L., Park, S.-B.: A big data-driven hybrid model for enhancing streaming service customer retention through churn prediction integrated with explainable AI. IEEE Access **12**, 69130–69150 (2024)
5. Chen, W.: Customer churn analysis for telecom operators based on SVM. In: Proceedings of the 3rd International Conference on Signal and Information Processing, Networking and Computers (ICSINC), vol. 473, pp. 327–332. Springer (2017)
6. Mishra, A., Reddy, U.S.: A comparative study of customer churn prediction in telecom industry using ensemble based classifiers. In: International Conference on Inventive Computing and Informatics (ICICI), pp. 721–725. IEEE (2017)
7. Petkovski, A.J., Stojkoska, B.L.R., Trivodaliev, K.V., Kalajdziski, S.A.: Analysis of churn prediction: a case study on telecommunication services in Macedonia. In: 2016 24th Telecommunications Forum (TELFOR), pp. 1–4. IEEE (2016)
8. Burez, J., Van den Poel, D.: Handling class imbalance in customer churn prediction. Expert Syst. Appl. **36**(3), 4626–4636 (2009)
9. Maldonado, S., López, J., Vairetti, C.: Profit-based churn prediction based on minimax probability machines. Eur. J. Oper. Res. **284**(1), 273–284 (2020)
10. De Bock, K.W., De Caigny, A.: Spline-rule ensemble classifiers with structured sparsity regularization for interpretable customer churn modeling. Decis. Support Syst. **150**, 113523 (2021)

11. Adhikary, D.D., Gupta, D.: Applying over 100 classifiers for churn prediction in telecom companies. Multimed. Tools Appl. **80**, 1–22 (2020)
12. Lemon, K.N., Verhoef, P.C.: Understanding customer experience throughout the customer journey. J Mark. **80**(6), 69–96 (2016)
13. Luo, X., Kumar, V.: Operational efficiency and customer retention in outsourced customer service operations. J. Mark. Res. **50**(2), 264–278 (2013)
14. Theodoridis, G., Tsadiras, A.: Applying machine learning techniques to predict and explain subscriber churn of an online drug information platform. Neural Comput. Appl. **34**(22), 19501–19514 (2022)
15. Tao, J., et al.: Explainable AI for cheating detection and churn prediction in online games. IEEE Trans. Games **15**(2), 242–251 (2023)
16. Guidotti, R., Ruggieri, S.: Ensemble of counterfactual explainers. In: DS 2021. LNCS, vol. 12986, pp. 358–368. Springer (2021)
17. Sharma, S., Henderson, J., Ghosh, J.: CERTIFAI: counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models (2019)
18. Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: unjustified counterfactual explanations (2019)
19. Samoilescu, R.-F., Van Looveren, A., Klaise, J.: Model-agnostic and scalable counterfactual explanations via reinforcement learning. CoRR, abs/2106.02597 (2021)
20. Bodria, F., Guidotti, R., Giannotti, F., Pedreschi, D.: Transparent latent space counterfactual explanations for tabular data. In: DSAA, pp. 1–10. IEEE (2022)
21. Ke, G., et al.:. LightGBM: a highly efficient gradient boosting decision tree. In: NIPS, pp. 3146–3154 (2017)
22. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: KDD, pp. 785–794. ACM (2016)
23. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
24. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: NIPS, pp. 4765–4774 (2017)