Causality-Driven Feature Engineering for Enhanced Telecommunications Churn Prediction

Gustavo F.V. de Oliveira
Inst. de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa
Florestal, Brazil
gustavo.viegas@ufv.br

Fabrício A. Silva
Inst. de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa
Florestal, Brazil
fabricio.asilva@ufv.br

Marcus H.S. Mendes
Inst. de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa
Florestal, Brazil
marcus.mendes@ufv.br

***Abstract*—Customer churn prediction is critical for telecommunication companies, but traditional machine learning models often struggle with the minority churn class and rely on correlational features. A possible solution to tackle this problem is to use causal-based feature engineering, which considers the causal relationship between the features and the churn decision. This study employs a causality framework to generate causal, tailored feature sets that enhance predictive accuracy, particularly for the minority churn class. Our findings indicate that feature sets engineered with causal insights significantly improve the discovery of underlying causal structures. In predictive tasks, these causally informed feature sets, especially when combined with oversampling, consistently outperformed baseline configurations in both test and cross-validation setups. These results demonstrate that integrating causal discovery into feature engineering is a potent strategy for developing more accurate, stable, and insightful churn prediction models.**

***Index Terms*—Causality, causal discovery, customer churn prediction, machine learning, causal feature engineering**

I. Introduction

Customer churn presents challenges across various industries, particularly within telecommunications [1]. Retaining existing customers is generally more cost-effective than acquiring new ones, and making precise customer churn predictions is crucial for businesses [2]. By predicting when a customer is likely to churn, companies can take proactive measures to enhance retention, improving long-term value (LTV) metrics [3].

Machine Learning (ML) is typically used to predict churn by analyzing historical data, including usage metrics and customer demographics [2]. However, these methods often use correlation rather than causation to uncover patterns and trends, leading to bias and incorrect predictions when dealing with unknown data [4].

Causality-based feature engineering offers a promising approach to enhance the effectiveness of churn prediction models by uncovering the underlying mechanisms that drive customer behavior. This study integrates causal discovery into feature engineering to develop improved feature sets for churn prediction using the *Causal-Nest* framework on a telecommunications churn dataset to identify features shaped by causal relationships [5]. The features are utilized to train predictive models, enabling a comparative analysis of their effectiveness, particularly in relation to the churn class, against models using traditional feature sets. The primary objective of this paper is to demonstrate that assessing causality in feature engineering for prediction problems can yield more accurate churn predictions, providing valuable insights into customer behavior and enabling more effective retention strategies [4].

The article is organized as follows: Section II presents the background regarding Churn Prediction, Causal Discovery, and Causality in Feature Engineering. Section III highlights the material and methods, the generated feature sets, and how the causal analysis was conducted. Section IV discusses experimental results regarding three ML models, test and cross-validation, and Section V summarizes the work and discussion.

II. Related Works

This section provides background on key areas relevant to our study: churn prediction, causal discovery, the role of causality in feature engineering, and the *Causal-Nest* framework used in our methodology.

A. Churn Prediction

Customer churn, which refers to customers stopping their engagement with a company, is a relevant issue, particularly in competitive markets like telecommunications [1]. Predicting churn enables businesses to implement targeted retention strategies, which are often more cost-effective than customer acquisition [2]. Traditional approaches typically involve ML models trained on historical customer data, including demographics, usage patterns, and service interaction history [3]. While effective to some extent, these models often capture correlations rather than underlying causal drivers, potentially leading to suboptimal interventions and poor generalization of new data distributions [4].

To predict customer churn effectively, it is crucial to focus on identifying class-1 customers (those who are likely to leave) more accurately than class-0 customers (those who are likely to stay). False negatives—where actual churners are misclassified as loyal customers—result in direct revenue loss and missed opportunities for retention [6]. In contrast, false positives merely incur the relatively minor cost of unnecessary retention campaigns [7]. This fundamental business reality

drives the focus toward metrics that emphasize the performance of the minority class (class-1, of churned customers) rather than overall classification accuracy.

In telecommunications churn prediction, recall typically takes precedence over precision due to the high customer lifetime value and competitive market dynamics [6]. The cost of losing a customer permanently far exceeds the expense of implementing retention strategies for non-churning customers, making comprehensive identification of at-risk customers the primary modeling objective. This emphasis on recall aligns with the business imperative to minimize false negatives, even if it comes at the expense of reduced precision.

B. Causal Discovery

Causal Discovery (CD) aims to uncover causal relationships from observational data, often represented as a Directed Acyclic Graph (DAG) where nodes represent variables and directed edges represent causal influences [8]. Key approaches include constraint-based methods that utilize conditional independence tests, score-based methods that search for the graph structure that optimizes a scoring criterion, and methods based on functional causal models. Each approach has different assumptions and strengths. Constraint-based methods are adept at identifying potential confounders, while score-based methods can be more efficient for larger datasets [9].

C. Causality in Feature Engineering

Feature engineering is a critical step in building effective predictive models. Traditional feature engineering often relies on domain knowledge or automated techniques that explore statistical relationships to inform feature selection. However, features derived from correlations only may not be robust or informative about the underlying mechanisms driving the outcome [4]. Incorporating causal insights into feature engineering aims to create features that represent causal pathways or mitigate confounding effects. This can potentially lead to more robust, interpretable, and generalizable models [4], [10]. For churn prediction, causal features may represent factors that directly influence a decision to leave, rather than merely correlating with it.

D. The Causal-Nest Framework

To systematically explore the impact of causal feature engineering, we employ the *Causal-Nest* framework [5]. *Causal-Nest* is designed to automate the execution and evaluation of multiple causal discovery algorithms on a given dataset. It generates candidate causal graphs, prioritizes them based on structural properties and consistency, and facilitates downstream tasks like feature selection or causal inference. Using *Causal-Nest*, we can efficiently generate causal graphs for different feature subsets and leverage these in our feature engineering process without requiring deep manual expertise in specific causal discovery algorithms.

This framework automates the execution and evaluation of numerous CD algorithms, allowing for a comparative analysis of the resulting graph structures. Typically, the process consists of a discovery phase, where causal graphs are generated by various algorithms, followed by an estimation and refutation phase that assesses the causal impact of the nodes within these graphs [5]. The overview of the framework pipeline is shown in Figure 1.

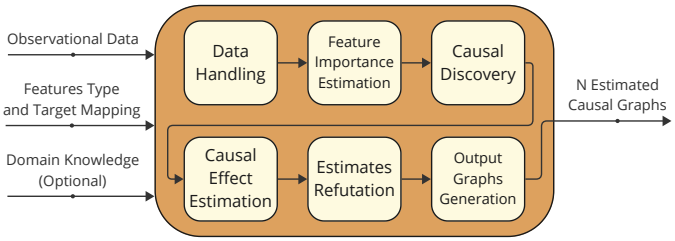

Fig. 1: *Causal-Nest* Pipeline, as originally presented in [5]

III. Methodology

This section outlines the methodology employed in our study, encompassing the dataset used, feature engineering strategies, including the integration of causal discovery via *Causal-Nest* [5], the metrics used for graph analysis, and the setup for churn prediction experiments.

A. Dataset Description

We utilize a publicly available customer churn dataset in an Iranian telecommunication company [11]. The dataset comprises 3150 instances and 14 features, including customer demographics, service usage patterns, billing information, and the target variable indicating whether a customer has churned. Hossain [2] utilizes the same dataset, which this study employs as a baseline for comparison and replication setup.

B. Dataset Configurations

This section describes the different feature sets utilized in the analysis. Each set represents a unique approach to feature engineering, aiming to uncover distinct patterns within the dataset for improved model performance. To ensure reproducibility, the source code utilized for the generation of each feature set is available in a dedicated GitHub repository [1].

1) Baseline: The baseline feature set consists of the original, raw features from the input dataset. This set serves as a control for comparison with more engineered feature sets, providing a fundamental understanding of model performance without advanced feature transformations, as presented by Hossain [2].

2) Causal: Through exhaustive experimentation with diverse feature combinations utilizing the causal inference pipeline of Causal-Nest, we have developed a promising causal feature set. This feature set outlines attributes that encapsulate critical elements of customer experience and behavior, which are hypothesized to have causal relationships with customer churn. Employing a causality-based framework enables the creation of diverse input data configurations for machine

[1] Available on github.com/gfviegas/itcp-causal-benchmark.

learning models, thereby lessening the dependence on domain specialists. This method highlights the effectiveness of causality as a valuable tool in feature engineering, enabling the generation of robust datasets without requiring expert input.

The causal feature set encompasses all baseline variables while introducing 24 novel engineered features designed to enhance predictive performance. Our feature engineering pipeline incorporates four distinct categories of causal features: customer behavior patterns, satisfaction indices, value segmentation, and direct causal constructs.

Customer Behavior Features capture usage volatility, engagement trends, and communication patterns that may indicate dissatisfaction or changing preferences. Customer Satisfaction Index synthesizes service quality indicators, including call failure rates, complaint frequency, usage satisfaction, and value perception, into a normalized composite score. Customer Value Segments partition customers into meaningful cohorts based on their monthly value and usage intensity, creating a 2×2 segmentation matrix that distinguishes between high and low value and usage patterns. The core Causal Features include eight key variables with hypothesized causal relationships to churn, listed below:

- *Complaint Severity*: combines complaints and call failures to quantify overall customer dissatisfaction.
- *Loyalty Index*: indicates customer allegiance through subscription length and usage frequency.
- *Service Quality*: reflects perceived service reliability by inversely relating to call failures per usage.
- *Value Perception*: assesses if a customer receives adequate value by comparing usage to the charged amount.
- *Communication Preference*: highlights shifts in customer interaction by balancing SMS and call frequency.
- *Network Breadth*: quantifies customer engagement with their network via distinct numbers to total calls ratio.
- *Customer Lifecycle*: approximates customer stage using age and subscription length.
- *Price Sensitivity*: evaluates customer responsiveness to pricing changes by comparing the charge amount to the customer value.

3) Baseline OS: The Baseline Oversampled (Baseline OS) feature set addresses potential class imbalance in the target variable by applying the Synthetic Minority Over-sampling Technique (SMOTE) to the baseline features. This process generates synthetic samples for the minority class, aiming to balance the class distribution and improve the model's ability to learn from the less frequent churn events. This set maintains the original feature space but adjusts the sample distribution.

4) Causal OS: Like Baseline Oversampled, the Causal Oversampled (Causal OS) feature set also employs SMOTE to mitigate class imbalance. However, it applies oversampling to the Causal feature set rather than the Baseline. This means that the synthetic samples are generated within the feature space of both the original and the causally engineered features. This approach aims to provide a balanced dataset while retaining the rich, causally-informed representations of customer behavior.

5) Causal Univariate: The Causal Univariate is a subset of the Causal feature set, originated by feature selection based on univariate statistical tests. It first generates the causal feature set and then selects the top 10 features with the strongest statistical relationship with the Churn target variable. This method aims to reduce dimensionality and retain only the most discriminative features from the causally engineered set, potentially improving model efficiency and interpretability by focusing on the most relevant individual predictors.

C. Causal Graph Generation and Analysis

To create causal graphs that illustrate potential relationships among the various feature sets outlined in Section III-B, we utilized the *Causal-Nest* framework. For this study, the *Causal-Nest* pipeline was configured to run on all feature sets presented in Section III-B. All features within these sets were treated as continuous variables to maximize the compatibility with various CD algorithms. The framework attempted to execute the following CD algorithms on each input feature set: Peter-Clark (PC), Grow-Shrink (GS), Concave Penalized Coordinate Descent with Reparametrization (CCDr), Incremental Association Markov Blanket (IAMB), Structural Agnostic Model (SAM), BIC Exact Search (BES), and Greedy Relaxation of the Sparsest Permutation (GRaSP).

A uniform timeout of 2 hours was imposed for each algorithm execution on each feature set to ensure fair comparison, acknowledging that algorithm complexity and runtime increase with the number of features (nodes). Due to this constraint, not all algorithms completed successfully for all feature sets. For the Baseline and Baseline OS feature sets, the following algorithms completed within the time limit: BES, CCDr, GRaSP, GS, IAMB, and PC. For the Causal, Causal OS, and Causal Univariate feature sets, only CCDr and PC completed successfully, indicating that the increased dimensionality led to timeouts for the other algorithms.

Following the CD step of the pipeline, *Causal-Nest* moved on to the phases of causal estimation and refutation for all identified graphs, completing this process without any timeouts. This pipeline was employed both for comparing feature sets through a causal lens and for generating the causal feature set in prior executions. In addition to the estimation results, we quantitatively assessed the generated graphs, particularly regarding their structural relevance to churn prediction, by calculating several standard graph metrics using network analysis libraries, as listed below:

- *Paths to Target (PTT):* Crucially for our hypothesis, we specifically counted the number of directed paths originating from other feature nodes and terminating at the *Churn* node. Generally, a greater number of identified paths indicates a higher probability that the discovery algorithm has revealed a significant causal connection. Subsequently, these paths are subjected to rigorous evaluation during the causal estimation and refutation, and therefore can be filtered by domain specialists.
- *Edges in Paths to Target:* We also counted the total number of unique edges participating in these directed

paths. This metric quantifies the direct and indirect structural connectivity leading to the target variable within the learned graph, which we hypothesize correlates with the predictive utility of the feature set.
- *Edges-to-Nodes Ratio*: The ratio of edges to nodes in a graph is an important measure for assessing the connectivity of generated graphs. A higher ratio can indicate richer interconnectivity, which may be beneficial in complex domains. However, there is no universal optimal value, as its significance varies depending on the dataset and context. This metric also enables the comparison of graphs generated by the same causal discovery algorithm across different feature sets, providing insight into how the composition of features influences the structure of learned causal relationships.

D. Experimental Setup

This study replicates the experimental framework established by Houssain [2], which presents sixteen different ML models. After a comprehensive assessment of all sixteen ML models, three models consistently demonstrated the most robust results for Class-1 churn prediction: the Decision Tree Classifier (DTC), the Gradient Boosting Machine Classifier (GBMC), and the Categorical Boosting Classifier (CBC). Our in-depth analysis, therefore, focuses on this select subset, employing the specified train-test split and k-fold cross-validation methodology. All models were initialized with a fixed pseudo-random seed of 42.

The DTC is recognized for its inherent interpretability, which is essential for deriving actionable insights into churn drivers by providing clear decision pathways [12]. It offers significant value for business stakeholders seeking to develop targeted retention strategies. Furthermore, the DTC serves as a fundamental benchmark for learners, offering a critical evaluation point against more complex ensemble methods. Conversely, the GBMC has advanced predictive capabilities as a highly effective gradient-boosting ensemble. The model is renowned for its sequential error correction mechanism and robust predictive performance, representing a leading approach in predictive analytics [13]. Complementing this, the CBC has the specialized ability to handle categorical features natively without extensive preprocessing. Its novel ordered boosting scheme also significantly enhances prediction stability and accuracy [14].

The three selected models represent a range of complexities, from interpretable foundational algorithms to advanced boosting techniques. This selection facilitates a thorough assessment of performance, interpretability, and practical applicability in real-world contexts of churn prediction.

For each combination of feature set and prediction model, we performed the following:

- **Data Splitting:** The data was split into training (80%) and testing (20%) sets using stratified sampling based on the churn variable to maintain class proportions.
- **Model Training:** Models were trained on the training set.
- **Evaluation Metrics:** Performance was evaluated on the test set using Accuracy, Precision, Recall, and F1 Score, calculated separately for the non-churn class (Class 0) and the churn class (Class 1). Our primary focus is on the metrics for Class 1 (Churn), particularly the F1 score, due to the typically imbalanced nature of churn datasets and the business importance of correctly identifying churning customers.
- **Cross-Validation:** To assess model generalization and stability, we also performed 10-fold stratified cross-validation on the training data for key feature sets, reporting the mean and standard deviation of the primary evaluation metrics (Accuracy, F1, Precision, Recall).

The experiments were performed on a 2023 MacBook Pro equipped with an Apple M3 Pro chip and 36 GB of RAM.

IV. Results and Discussion

This section details the experimental outcomes, evaluating the performance of selected machine learning models, CatBoost (CBC), Gradient Boosting (GBMC), and Decision Tree (DTC), across five distinct feature sets: Baseline, Baseline Oversampled (Baseline OS), Causal, Causal Oversampled (Causal OS), and Causal Univariate.

A. Causal Graphs Metrics

This section presents an analysis of the causal graphs discovered by various causal discovery (CD) algorithms for each feature set, as summarized in Table I. The focus is on metrics that reflect the causal insights gained: Paths to Target (PTT), Edges in PTT, and Edges-to-Nodes Ratio.

CD Algorithm	Feature Set	PTT	Edges in PTT	Edges-to-Nodes Ratio
PC	Baseline	1	1	1.57
	Baseline OS	5	8	1.71
	Causal	**40**	**172**	1.47
	Causal OS	8	18	**1.89**
CCDr	Baseline	1	1	1.79
	Baseline OS	7	12	**4.00**
	Causal	**882**	**6647**	3.13
	Causal OS	0	0	0.74
	Causal Uni.	345	1509	**4.00**
BES	Baseline	**600**	**3162**	3.79
	Baseline OS	14	40	**4.36**
	Causal Uni.	24	62	3.09
GRaSP	Baseline	**1**	**1**	2.79
	Baseline OS	**1**	**1**	**3.57**
	Causal Uni.	**1**	**1**	2.55
IAMB	Baseline	1	1	**1.64**
	Baseline OS	**2**	**2**	1.57
GS	Baseline	0	0	**1.64**
	Baseline OS	**2**	**2**	1.57

TABLE I: Causal graphs metrics for each feature set

The causal graph discovery process aimed to identify direct and indirect causal pathways to the target variable *Churn*. The metrics in Table I reveal significant differences in the structure and density of the discovered causal graphs across various CD algorithms and feature sets.

A notable observation is the behavior of algorithms like GRaSP, IAMB, and GS. While these algorithms were constrained by time limits for the Causal and Causal OS feature sets, where they could not produce results, their successful executions for other feature sets generally resulted in graphs with a very low number of Paths to Target (PTT) and a minimal count of Edges in PTT (typically 1 or 2). This suggests that these algorithms, under the given conditions, struggle to identify complex or extensive causal relationships leading to the target, indicating that the graphs generated by these algorithms may offer limited causal insights, regardless of the feature set.

The PC and CCDr algorithms exhibited a more comprehensive discovery of causal structures compared to the baseline. The graph generated by PC on the Causal feature set achieved an impressive 40 PTT with 172 Edges in PTT, a substantial increase over the Baseline (1 PTT, 1 Edge) and Baseline OS (5 PTT, 8 Edges). This suggests that the Causal feature set enhances the potential of the PC algorithm to uncover a more intricate causal network toward the target. Additionally, the Causal OS for PC shows improvements in both PTT and Edges in PTT over the baseline. The CCDr algorithm further enhances findings, with its Causal feature set yielding a remarkable 882 PTT of 6,647 Edges in PTT, outpacing all other combinations. This indicates that the Causal feature set effectively uncovers a complex causal structure relevant to the target. Although Causal OS for CCDr yielded no paths, Causal Univariate (345 PTT, 1,509 Edges in PTT) still demonstrated a strong capacity to reveal causal pathways, highlighting the advantages of using causally informed features.

Regarding the BES algorithm, the Baseline feature set successfully generated a substantial number of paths to the target, characterized by a high volume of edges. However, the nodes included in these paths lacked causal strength, as will be demonstrated in the estimates presented in the following section. Overall, the BES algorithm lacked the power to unveil the causal connections in this particular setup, regardless of the feature set.

The Edges to Nodes Ratio indicates overall graph density, where the Causal and Causal OS feature sets for PC and CCDr often demonstrate competitive or superior ratios, particularly in targeting pathways. Notably, the Causal feature set for CCDr achieves an Edges-to-Nodes Ratio of 3.13, along with a substantial number of target-specific paths. Conversely, while other graphs may display similar ratios, their edges are not necessarily aligned with causal paths to the outcome.

The examination of total edges within causal pathways leading to the target reinforces these findings. The graphs generated from the Causal feature set, predominantly influenced by CCDr, produced 6,819 edges in causal paths—effectively doubling the total from the Baseline feature set, despite the analysis involving four fewer graphs due to time limitations. This highlights the significant benefits of utilizing causally engineered features to uncover a more accurate causal structure for churn prediction.

B. Causal Estimations Metrics

This section presents an aggregated view of the causal estimation values derived from the successfully executed Causal Discovery (CD) algorithms for each feature set. The causal estimation results provide quantitative insights into the direct and indirect influence of features on the churn outcome, as identified by the respective causal discovery algorithms. Table II summarizes the number of causal estimates where their absolute value is positive and the mean absolute value for each CD algorithm and feature set.

CD Algorithm	Feature Set	Number of Estimates	Mean Abs. Estimates
PC	Baseline	1	0.73
	Baseline OS	5	0.10
	Causal	**17**	**2.27**
	Causal OS	6	0.23
	Causal Uni.	0	-
CCDr	Baseline	1	0.73
	Baseline OS	3	0.18
	Causal	**22**	**1.80**
	Causal OS	0	-
	Causal Uni.	0	-
BES	Baseline	11	0.04
	Baseline OS	**5**	**0.12**
	Causal Uni.	0	-
GRaSP	Baseline	1	**0.73**
	Baseline OS	1	0.54
	Causal Uni.	0	-
IAMB	Baseline	1	0.73
	Baseline OS	**2**	0.26
GS	Baseline	0	-
	Baseline OS	**22**	**0.26**

TABLE II: Causal estimations metrics for each feature set

The data presented in the Table II demonstrates that the PC algorithm exhibits a distinct advantage when employing the Causal feature set. This configuration enables the identification of 17 significant causal estimates, accompanied by a mean absolute estimate of 2.27. These results significantly surpass the number of substantial causal estimates identified by both the Baseline and Baseline OS sets for the PC algorithm, which report only 1 and 5, respectively. This observation underscores the critical role that integrating causal insights into feature engineering plays in enhancing the identification of direct causal effects.

Similarly, the CCDr algorithm shows exceptional performance with the Causal feature set, yielding 22 significant causal estimates and a mean absolute estimate of 1.80. This performance not only reflects a high volume of identified causal relationships but also indicates a considerable strength of these relationships on average. In contrast, the Baseline and Baseline OS feature sets, while identifying some causal estimates (1 and 3, respectively), demonstrate significantly lower mean absolute effect magnitudes of 0.73 and 0.18. Such findings strongly suggest that the Causal feature set provides CCDr with the essential information needed to uncover a larger quantity of consequential and substantively stronger causal links.

C. Performance Metrics

This section details the performance of the Categorical Boosting Classifier (CBC), Gradient Boosting Machine Classifier (GBMC), and Decision Tree Classifier (DTC) models on various feature sets, evaluated using a single 80/20 train-test split. The Class-1 F1 Score, Precision, and Recall are presented in Figure 2, Figure 3, and Figure 4, respectively.

The F1 Score, which balances precision and recall, is a relevant metric for imbalanced classification tasks, such as churn prediction. Figure 2 shows that for the CBC model, the Causal OS feature set achieved the highest F1 Score of 90.55%, closely followed by the Baseline feature set with a score of 90.36%. This suggests that combining causally-informed features with oversampling effectively optimizes CBC balanced performance. The GBMC model also attained its highest F1 Score of 89.30% with the Causal OS feature set. The Causal feature set (without oversampling) also performed strongly for GBMC, yielding a F1 Score of 88.21%. This indicates that both causally engineered feature sets have the potential to enhance the predictive efficacy of GBMC. In contrast, the DTC consistently exhibited lower F1 Scores across all configurations, with its best performance reaching 82.47% using the Causal feature set. This highlights the inherent limitations of a single decision tree in achieving the same level of balanced performance as ensemble methods. However, the DTC provides valuable insights when inspecting its rules.

A trend is observed across the models: the interplay between precision and recall, particularly when oversampling is applied. Recall, as shown in Figure 4, values for oversampled feature sets, Baseline OS, and Causal OS generally led to higher Class-1 Recall values for all models. For instance, GBMC with Causal OS achieved a high recall value of 96.97%, and CBC with Baseline OS reached 92.93%. This improvement suggests an enhanced ability to correctly identify true churners, which is often a primary objective in churn prediction. Precision values, as shown in Figure 3, were slightly lower compared to not oversampling. For example, the precision of GBMC dropped from 87.50% with Baseline to 69.23% with Baseline OS, even as its recall significantly increased. This illustrates the typical trade-off in imbalanced classification: increasing the true positive rate often leads to a higher rate of false positives. Notably, the Causal feature set yielded the highest precision for both CBC (94.32%) and GBMC (89.58%), suggesting that causally informed features can contribute to more accurate positive predictions.

The introduction of causally-informed features generally led to improved performance for the churn class compared to the baseline. Notably, the Causal feature set yielded better F1 scores for both GBMC and DTC compared to the Baseline set. While CBC performed exceptionally well on the Baseline, its performance remained strong with the Causal and Causal OS sets. The Causal OS feature set provided the highest F1 score for CBC and GBMC among the tested configurations, demonstrating the combined benefit of causal features and addressing class imbalance. The Causal Univariate set, representing a feature selection approach on the causal features, offered competitive performance, particularly for DTC, suggesting that a subset of causal features retains significant predictive power.

D. Cross Validation

The robustness of the models was assessed using 10-fold stratified cross-validation. The oversampling of the feature sets was conducted solely on the training dataset, ensuring that the validation fold remained unaffected by the oversampling techniques. This approach maintains the integrity of the validation process by providing an unbiased assessment of the model. The mean and standard deviation across folds are presented in Table III.

		(Mean ± Std)		
	Feature Set	**Class-1 F1 Score**	**Class-1 Precision**	**Class-1 Recall**
CBC	Baseline	86.88 ± 2.40	90.35 ± 2.76	83.86 ± 4.45
	Baseline OS	88.24 ± 2.87	84.75 ± 2.78	**92.17 ± 4.55**
	Causal	87.20 ± 2.82	**91.33 ± 3.52**	83.60 ± 4.35
	Causal OS	**88.86 ± 3.10**	85.98 ± 3.09	**92.17 ± 5.21**
	Causal Uni.	83.24 ± 4.18	85.76 ± 4.57	81.08 ± 5.57
DTC	Baseline	79.60 ± 5.10	**81.94 ± 5.14**	77.53 ± 5.93
	Baseline OS	80.42 ± 4.11	78.17 ± 5.96	**83.08 ± 4.07**
	Causal	79.48 ± 5.05	80.95 ± 5.77	78.31 ± 6.30
	Causal OS	**80.45 ± 5.16**	78.82 ± 6.30	82.36 ± 5.43
	Causal Uni.	76.55 ± 3.47	74.99 ± 3.84	78.85 ± 7.85
GBMC	Baseline	80.89 ± 3.55	86.65 ± 4.40	76.03 ± 4.55
	Baseline OS	79.27 ± 3.79	70.37 ± 4.40	90.92 ± 4.23
	Causal	83.27 ± 1.91	**87.32 ± 3.95**	79.81 ± 3.47
	Causal OS	**84.90 ± 3.94**	79.29 ± 4.41	**91.69 ± 5.94**
	Causal Uni.	79.55 ± 4.23	83.06 ± 5.78	76.53 ± 4.62

TABLE III: Mean and standard deviation metrics of 10-fold cross validation achieved by different feature sets

The findings reinforce the trends observed in the single split evaluation, with the Causal OS feature set consistently yielding higher average F1 scores for the churn class across the various folds. The implementation of oversampling techniques has demonstrated significant importance in generating robust recall values of the churn class. Both the Baseline OS and Causal OS sets produced superior recall values across all evaluated models.

Another significant observation across all models and feature sets is the relatively high standard deviations observed for all reported metrics. This suggests a notable variability in model performance across the 10 cross-validation folds. Such variability may arise from the inherent complexity and potential noise of the churn prediction dataset. These are impacted by the small size of the dataset and the high imbalance between the churn and non-churn classes.

The CBC consistently showcased the strongest overall performance, achieving the highest Class-1 F1 Scores across various feature sets. Notably, the Causal OS feature set produced the highest F1 Score of 88.86% ± 3.10. This outcome suggests a synergistic effect deriving from the advantages of oversampling combined with the potentially richer information provided by causally engineered features. Additionally, the effective handling of categorical features of CBC, along with

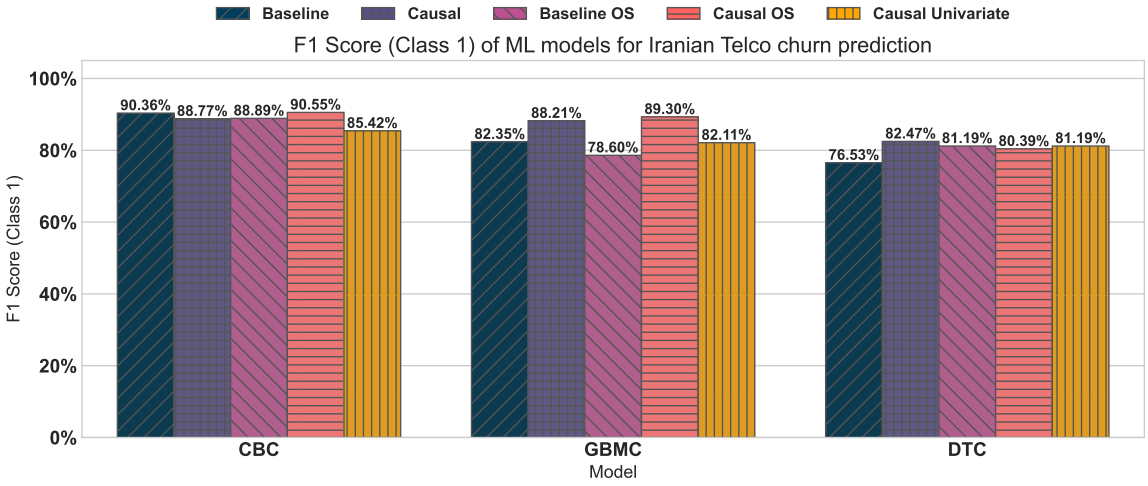

Fig. 2: F1 Score metrics for the churn class of ML models for each feature set

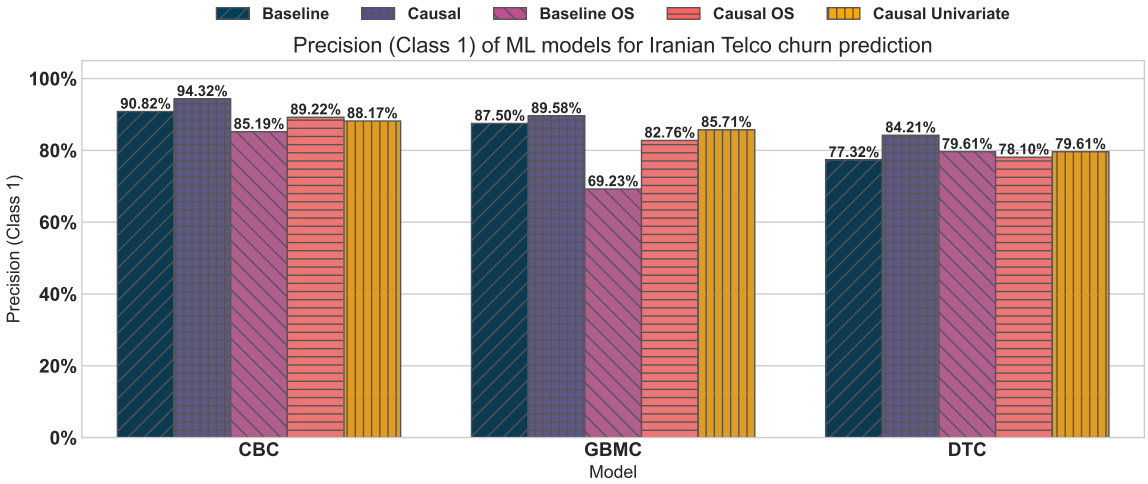

Fig. 3: Precision metrics for the churn class of ML models for each feature set

its ordered boosting scheme, likely contributes to its superior performance and stability when working with oversampled data.

The GBMC demonstrated notable predictive capabilities, particularly when utilizing the Causal OS feature set, which achieved an F1 Score of 84.90% $\pm$ 3.94. Similar to the CBC, the GBMC experienced a significant enhancement in Class-1 Recall with oversampled data, aligning with the expected behavior of boosting algorithms when exposed to a higher number of minority class samples. While the Causal OS maximized its F1-score, the Causal feature set achieved its highest Class-1 Precision (91.33% $\pm$ 3.52 for CBC and 87.32% $\pm$ 3.95 for GBMC). This indicates that causal features alone can facilitate more accurate identification of churners, particularly when the model is not primarily optimized for recall.

The DTC demonstrated lower F1 Scores (76.55% to 80.45%) and higher standard deviations than ensemble methods, highlighting the limitations of using a single decision tree on complex datasets prone to variance. While oversampling improved Class-1 Recall, the F1 Score increase was minimal, indicating that balancing class distribution alone doesn't resolve the instability of a standalone decision tree. However, incorporating causally engineered features led to improved metrics compared to baseline counterparts.

The Causal Univariate feature set consistently underperformed across all models, indicating it may lack predictive power or restrict the model's ability to capture diverse patterns. In contrast, the Causal OS feature set proved most effective for the ensemble models (CBC and GBMC), achieving the highest F1 Scores due to improved minority class representation through oversampling and more informative causally-derived features.

V. Conclusion

This paper investigates the application of causal discovery, facilitated by the *Causal-Nest* framework, to enhance feature engineering for predicting customer churn in the telecommunications industry. By generating causal graphs from ob-

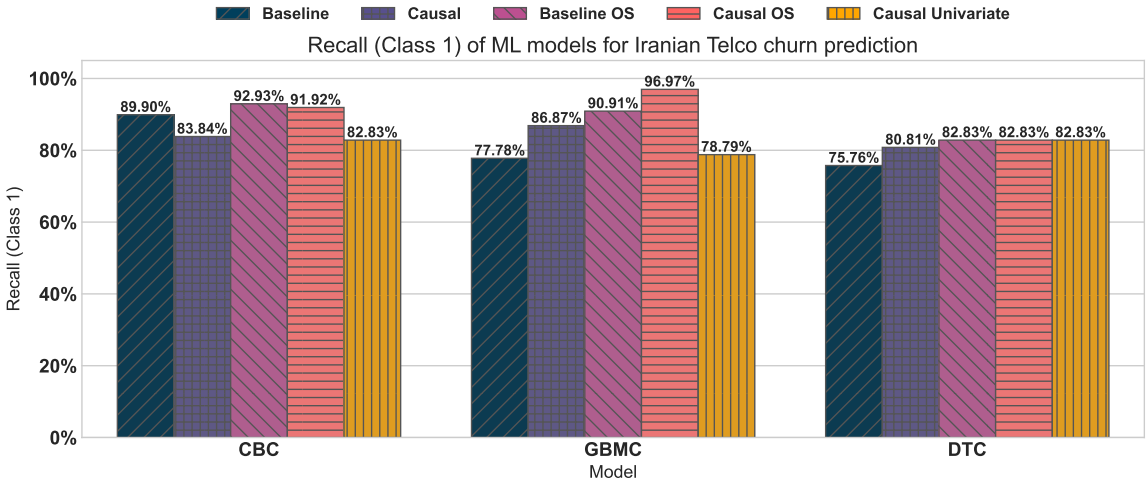

Fig. 4: Recall metrics for the churn class of ML models for each feature set

servational data and deriving feature sets informed by these structures, we aimed to improve the predictive performance, particularly for the minority churn class, compared to the baseline raw data.

Our experimental results show that incorporating causal insights into feature engineering significantly enhances the analysis of churn. Causally-informed feature sets, particularly Causal and Causal OS, improved the identification of causal structures, enabling algorithms like CCDr and PC to discover more paths and edges linked to churn drivers. The causal estimation analysis further revealed stronger statistically significant relationships with these feature sets.

For predictive performance, ensemble models such as the Categorical Boosting Classifier (CBC) and Gradient Boosting Machine Classifier (GBMC), alongside a Decision Tree Classifier (DTC), were tested. The Causal OS feature set yielded the highest F1 Scores, demonstrating the effectiveness of causal feature engineering and class imbalance handling through oversampling. While this improved Class-1 Recall, it sometimes reduced Class-1 Precision, but the overall F1 Score suggests a good balance. Although the Causal Univariate feature set underperformed, it holds promise for future dimensionality reduction.

Given the modest dataset size (3,150 rows) and class imbalance, caution is needed in generalizing these findings. The complex causal graphs generated by CD algorithms may indicate overfitting or noise. Future work should test this methodology on other datasets and integrate causal insights into prescriptive analytics to inform actionable strategies for reducing churn.

References

[1] A. Manzoor, M. Atif Qureshi, E. Kidney, and L. Longo, "A review on machine learning methods for customer churn prediction and recommendations for business practitioners," *IEEE Access*, vol. 12, 2024, DOI: 10.1109/ACCESS.2024.3402092.

[2] M. R. Hossain, "Predicting customer churn in telecommunications with machine learning models," *Asian Journal of Research in Computer Science*, vol. 18, no. 1, 2025, DOI: 10.9734/ajrcos/2025/v18i1548.

[3] A. Tamaddoni Jahromi, S. Stakhovych, and M. Ewing, "Managing B2B customer churn, retention and profitability," *Industrial Marketing Management*, vol. 43, no. 7, pp. 1258–1268, 2014, DOI: 10.1016/j.indmarman.2014.06.016.

[4] D. H. Rudd, H. Huo, and G. Xu, "Causal analysis of customer churn using deep learning," in *2021 International Conference on Digital Society and Intelligent Systems (DSInS)*, 2021, pp. 319–324, DOI: 10.1109/dsins54396.2021.9670561.

[5] G. F. de Oliveira, F. A. Silva, and M. H. Mendes, "A framework for practical causal discovery from observational data," in *Proceedings of the 2nd Workshop on Causal Discovery - CaDis*, Universidad de la República, Montevideo, Uruguay, Nov. 2024. [Online]. Available: https://amexcomp.mx/media/publicaciones/CaDis_2024_Proceedings.pdf

[6] S. Ouf, K. T. Mahmoud, and M. A. Abdel-Fattah, "A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry," *Journal of Big Data*, vol. 11, no. 1, May 2024, DOI: 10.1186/s40537-024-00922-9.

[7] Y. Zhou, W. Chen, X. Sun, and D. Yang, "Early warning of telecom enterprise customer churn based on ensemble learning," *PLOS ONE*, vol. 18, no. 10, Oct. 2023, DOI: 10.1371/journal.pone.0292466.

[8] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT Press, 2000, DOI: 10.1007/978-1-4612-2748-9.

[9] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. Volume 10 - 2019, 2019, DOI: 10.3389/fgene.2019.00524.

[10] K. Yu, X. Guo, L. Liu *et al.*, "Causality-based feature selection: Methods and evaluations," *ACM Comput. Surv.*, vol. 53, no. 5, Sep. 2020, DOI: 10.1145/3409382.

[11] "Iranian Churn," UCI Machine Learning Repository, 2020, DOI: 10.24432/C5JW3Z.

[12] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, p. 36–43, Sep. 2018, DOI: 10.1145/3233231.

[13] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, vol. 29, no. 5, 2001, DOI: 10.1214/aos/1013203451.

[14] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018.