EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

Explainable AI for Customer Churn Prediction in the Energy Sector Using Ensemble Machine Learning Models

Abdullah Al Mamun[1], Md Fazla Saim Tanoor[1], Abdul Kadar Muhammad Masum[2], Touhid Bhuiyan[3], Md. Maruf Hassan[2,*]

[1]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering Southeast University, Dhaka, Bangladesh
[3]School of IT, Washington University of Science and Technology, VA, USA

Abstract

Customer churn prediction is crucial for energy providers to preserve revenue and market share in a competitive setting. This research explores implementing explainable AI (XAI) in customer churn prediction with machine learning algorithms as well as interpretability methods. This research utilizes a dataset formed by combining client and price information from Kaggle's PowerCo dataset, making use of the 'id' column as a key. During the data preprocessing section, the data involves extensive preparation, such as dropping unnecessary columns, deleting duplicates, encoding category features, capping outlier values, and imputing missing values. The class imbalance problem was alleviated through making use of the Synthetic Minority Oversampling Technique (SMOTE) to enable robust training of models. The machine learning algorithms Random Forest, XGBoost, and LightGBM were built, implemented, and benchmarked with principal performance metrics like accuracy, precision, recall, F1-score, as well as ROC-AUC. Among the evaluated ensemble models, XGBoost demonstrated the most balanced trade-off between predictive accuracy and interpretability when combined with SHAP and LIME. SHAP (SHapley Additive exPlanations) as well as LIME (Local Interpretable Model-agnostic Explanations) are used for providing global as well as local interpretability to identify key drivers of churn. This proposed approach exhibits potential to provide accurate predictions as well as salient insights, enabling energy providers to design targeted retention strategies.

Received on 29 July 2025; accepted on 11 March 2026; published on 07 April 2026

Keywords: Customer Churn Prediction, Explainable AI (XAI), Machine Learning, Random Forest, XGBoost, LightGBM, SMOTE, SHAP, LIME, Performance Metrics, ROC-AUC, PR-AUC.

doi:10.4108/airo.9813

1. Introduction

Customer churn prediction is now an important factor for energy companies, as keeping existing customers is cheaper than finding new customers [4]. Having an accurate prediction capacity about who is going to churn is crucial in competitive markets as energy providers can customize their retention initiatives and allocate assets more effectively to secure customer relationships and maintain revenue streams [35]. Churn prediction is difficult, as customer behavior is complex and is influenced by various factors behind their intention to switch energy providers, including pricing schemes, service quality, contract duration and terms, and broader industry conditions [1], [3], [18], [19].

New breakthroughs in machine learning have contributed to more accurate prediction abilities in churn detection through complex models that have the capability to capture subtle patterns and also interact among features [36]. However, most high-accuracy models, popularly described as "black-box" models, are opaque and do not allow explanation, making it problematic to implement them in practice [37]. Energy domain experts and decision-makers require

*Corresponding author. Email: ancssf@gmail.com

explainable and interpretable predictions to believe in model outputs and implement effective decisions [21], [38].

Explainable Artificial Intelligence (XAI) overcomes this challenge by offering ways to explain complex model predictions [39]. Two common XAI tools that provide global and local explanation perspectives are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) [40]. These methodologies allow for finding out the most impactful features behind predictions and interpreting single decisions [41], and thus making outputs explainable to stakeholders [6], [7], [42].

This study examines customer churn prediction using XAI in the energy industry through the evaluation and comparison of three strong machine learning algorithms: Random Forest, XGBoost, and LightGBM. Drawing upon a solid dataset on Kaggle's PowerCo [29], which integrates customer demographics and pricing data, the research preprocesses data through exhaustive cleansing, feature encoding, outlier treatment, and synthetic oversampling (SMOTE) to correct for class imbalance.

The performance of all the models is measured using various parameters like accuracy, precision, recall, F1-score, ROC-AUC, and PR-AUC. The top-scoring model, XGBoost, is explained further with SHAP for global feature evaluations and LIME for local interpretations to support predictions and give actionable results. The aim is to prove that the incorporation of XAI into churn prediction models makes them more trustworthy and usable for energy companies so that they are able to devise better customer retention strategies.

Unlike churn prediction in domains like telecommunications or banking, customer attrition in the energy sector is deeply affected by context-specific variables such as dynamic pricing structures (mid-peak and off-peak rates), regulatory caps on tariffs, seasonal demand fluctuations, and fixed-term contractual obligations. These produce quite distinct behavioral and financial patterns that the conventional churn frameworks do not effectively capture. This called for re-engineering of the preprocessing entailment with this domain, such as pricing feature normalisation and contract-based segmentation, and interpreting the results in view of sectoral constraints. This specialization underlines the novelty of our approach beyond a simple application of existing models.

In addition, the proposed framework corresponds to transformations currently taking place within the energy sector. As utilities continue to decarbonize and grow renewable offerings [31], both customer engagement and retention are increasingly crucial to meet sustainability and policy goals. Transparent churn prediction models support not only financial stability but can also help providers design retention strategies that drive participation in renewable and time-of-use programs [30]. By coupling interpretability with emergent green-market dynamics, the study contributes toward a wider agenda of trustworthy and sustainable AI in energy analytics.

In summary, we propose an approach to customer churn prediction in the energy sector using explainable AI techniques. This paper consists of the following contributions:

1. Combined explainable AI methods, SHAP and LIME, with ensemble machine learning models to forecast customer churn for the energy industry.
2. Built an overall data preprocessing pipeline consisting of data cleansing, categorical encoding, outlier treatment, and class balance via SMOTE to ensure model resilience.
3. Evaluated three state-of-the-art ensemble algorithms, Random Forest, XGBoost, and LightGBM, using a number of different evaluation measures such as accuracy, precision, recall, and F1-score.
4. Delivered deep interpretability analyses that identify important drivers of customer churn and provide energy providers with actionable insights.
5. Demonstrated that our model, XGBoost, outperforms existing models based on not only model accuracy but also interpretability and therefore practicality of deployment.

The sections to follow within this paper present a concise overview of significant literature relating to customer churn modeling and explainable artificial intelligence. Section 3 describes the dataset, preprocessing steps, training strategy, and evaluation metrics. Section 4 introduces the ensemble learning algorithms used in this study. Section 5 reports experimental results, comparative performance, and interpretability analysis using SHAP and LIME. Section 6 outlines future research directions, and Section 7 concludes the paper by highlighting major findings and implications.

2. Related works

In the telecoms, financial, and retail industries, where behavioral data is usually high-frequency and transactional right away, a lot of study has been done on how to forecast client attrition. But the energy industry has its own structural challenges, like consumption cycles that change with the weather, price plans that are set by policy, and contracts that keep customers. These issues alter both the distributions of characteristics and the requirements for interpretability. Even while ensemble learning and XAI techniques have been useful in other areas, they haven't been employed much

in energy analytics yet. They need to be changed for the energy field so that they can provide you valuable information. If accurately predicted by proper modeling, companies can minimize attrition through targeted marketing to high-risk customers [1], [11].

Early prediction models for churn were based largely on classical statistical methods using logistic regression and survival analysis, which are highly interpretable but less capable of modeling complex associations in high-dimensional datasets [12]. Since the advent of machine learning, more advanced algorithms including decision trees, random forests, support vector machines (SVM), and boosting algorithms have been utilized for churn prediction with the added advantage of fitting nonlinear and interaction terms [13], [5].

Among them, ensemble approaches such as Random Forest and gradient boosting approaches including XGBoost and LightGBM, have become state-of-the-art as a result of their reliability, scalability, and better generalizability in modeling heterogeneous data [14], [15]). These approaches take advantage of multiple weak learning machines to minimize variance and bias and hence are suited to imbalanced class churn datasets and noisy attributes [32]; [11].

A major challenge in modeling churning is handling imbalanced datasets with significantly smaller numbers of churners as compared to non-churners, causing model training to be biased towards the majority class. Oversampling using synthetic oversampling algorithms like SMOTE (Synthetic Minority Oversampling Technique) is common in producing synthetic minority class instances in order to balance training and enhance model sensitivity [2], [17]).

Moreover, churn datasets tend to have missing values, categorical variables, and outliers, and hence need special preprocessing processes such as imputation, encoding, and outlier capping to maintain model stability and generalizability [28]. Feature selection and feature engineering based on domain expertise and data exploration also affect model performance and interpretability [12].

Though machine learning algorithms provide greater accuracy, their intricacy generally hinders interpretability, an aspect necessary for deployment in customer-focused industries in practice. Explainable AI (XAI) is a response to fill in the gap by proposing tools for interpretation and explanation of model decisions [21].

SHAP (SHapley Additive exPlanations) utilizes game theory concepts to attribute each feature a contribution score for specific predictions and provides global feature importances as well as local explanations for individual predictions [6]. The use of SHAP values enables domain experts to better realize feature effects and interdependencies and promotes trust and actionability.

LIME approximates a local region around a prediction using an interpretable model, often linear, to provide an explanation for the prediction in a human-interpretable form [7]. LIME supplements SHAP as it offers local explanations for particular instances and is likely to uncover heterogeneity in customer behavior [8].

Beyond attribution-based methods, counterfactual explanation frameworks can translate predictions into actionable "what-if" prescriptions (i.e., minimal changes to features that flip the outcome), and privacy-preserving training via federated learning enables cross-utility modeling without sharing raw customer data. We position our XGBoost-centered SHAP/LIME design as a strong baseline and outline how these emerging approaches could extend its operational impact in Sections 5.4 and 6 [9], [10].

A number of studies have used SHAP and LIME to perform churn analysis. For instance, K. Peng et al. [22] illustrated how SHAP added explanation to XGBoost models for telecommunication churn prediction. Likewise, Tiwari et al. [23] highlighted model interpretability in high-risk fields and illustrated how LIME generates actionable information for single decisions [25]. Recent energy-sector work likewise demonstrates the utility of SHAP for interpretable forecasting and feature selection, reinforcing the case for explainable modeling in energy analytics [24]. In addition, EAI Endorsed Transactions on Energy Web has reported machine learning applications in energy analytics such as intelligent energy storage management and wind energy prediction [26], [27].

Churn forecasting in the energy industry is a rather nascent field with unique challenges as a result of varied customer profiles, contract terms and conditions, and consumption patterns. Energy companies experience high competition due to emerging entrants in their markets and regulatory pressures, making reducing churn a desirable goal [20].

Past studies on energy churn prediction have utilized customer consumption information, contract information, and price data [16]. Yet, almost all center on model accuracy and overlook explainability, hindering real-life application among energy firms.

Integrating XAI approaches such as SHAP and LIME into energy churn models presents a viable path towards identifying influential drivers of churn contract duration, pricing factors, and consumption habits to guide targeted retention initiatives and improve stakeholders' confidence [11].

In spite of ongoing progress, existing literature is short on integrated approaches that leverage strong machine learning models together with explainable artificial intelligence approaches specific to energy

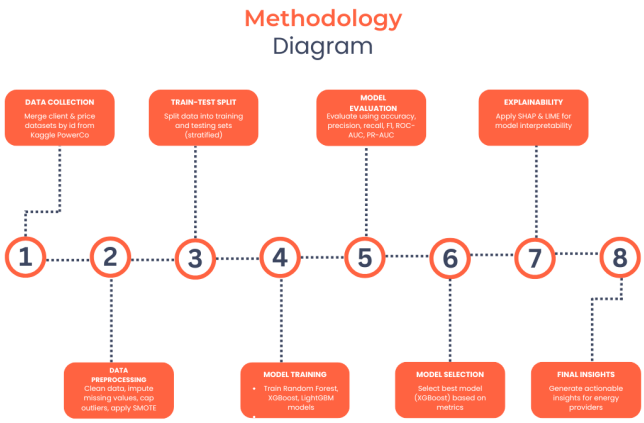

Figure 1. Process Flow Diagram

industry churn. Several studies concentrate on prediction performance alone or achieve superficial interpretability, lacking domain-based support for feature contributions [33].

In addition to that, few studies compare various machine learning models with XAI integration on a common energy churn dataset. There is very little work that examines the synergy between global explainability (SHAP) and local instance explanation (LIME) to provide actionable and dependable insights.

This research bridges those gaps through a holistic method that integrates data pre-processing, model evaluation (Random Forest, XGBoost, and LightGBM), and twinned XAI approaches (SHAP and LIME) on an energy customer churn database. This illustrates the utility of explainability in model prediction validation and identifying actionable factors to enable real-world applications in energy firms.

3. Materials and Methods

This section provides data information and evaluation metrics for this study.

3.1. Methodological Pipeline and Dataset

The primary goal of this study is to find the best and robust machine learning model to predict customer churn in the energy sector. For that purpose, we utilized various machine learning techniques such as Random Forest, XGBoost, and LightGBM. For both the training and evaluation of each model, the model's performance was measured based on essential evaluation metrics in order to pick the best-performing model.

The proposed method is based on the acquisition of Kaggle's PowerCo dataset, which includes customer attributes and pricing data. After data collection, a preprocessing step is applied, which includes data cleaning, missing value replacement, encoding of categorical features, outlier handling, and balancing of the classes using SMOTE.

To ensure reproducibility and optimal performance, hyperparameter tuning was performed within the cross-validation framework using grid search. For XGBoost, the best configuration included a learning rate of 0.1, a maximum depth of 6, 200 estimators, and a subsample ratio of 0.8. For Random Forest, 300 trees with a maximum depth of 10 and Gini impurity criterion were used. LightGBM was tuned with a learning rate of 0.05, 250 estimators, and a feature fraction of 0.8. Default parameters were retained where tuning did not improve the validation F1-score.

Lastly, the best model (referred to as XGBoost in this work) is analyzed using explainable AI techniques like SHAP and LIME. These methods offer global and local interpretability, aiding users to identify key churn drivers and trust their predictions while providing actionable insights for decision-making in the energy sector (Fig. 1).

The data used for this research were combined into one dataset using data acquired from Kaggle's PowerCo repository [29]. The datasets contain both client-specific data, including demographics and contract data, and pricing data associated with different energy plans. The datasets were merged based on the common ID column to match clients' data with their respective pricing data to create an integrated dataset that is utilized for churn prediction modeling. The combined dataset is made up of 12,241 instances; 90.9% (11,123 instances) comprise non-churned customers, and 9.1% (1,118 instances) comprise churned customers. Such class imbalance is typical in real-world applications in the energy industry because most customers stick with their services, while a smaller percentage drop out.

Every case includes a mix of numerical and categorical attributes representing customer demographics and contract features, as well as energy consumption habits and granular price features. Prominent attributes comprise contract intervals, contract modification intervals, different prices in terms of mid-peak and off-peak periods, and calculated margin figures assigned to electricity consumption. These attributes reveal information about customer behavior as well as economic conditions driving churn (Fig. 2).

3.2. Data Preprocessing

Proper data preprocessing is essential to make machine learning models accurate and perform well. The raw data is normally full of inconsistencies, missing records, categorical variables, and outliers that must be dealt with prior to training. The steps adopted below were used to preprocess data for churn prediction models.

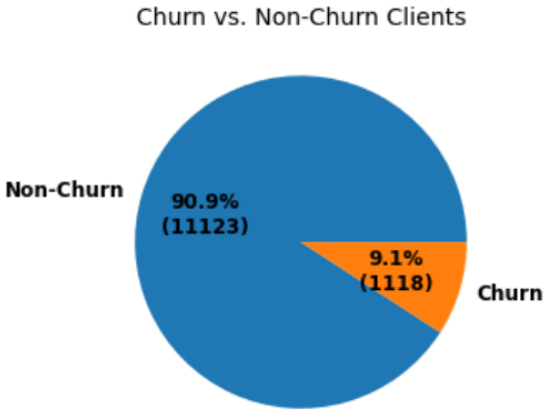

Figure 2. PowerCo Customer Churn Datasets

Handling Missing Data Missing values within the dataset may result in biased or unreliable model output if left unaddressed. Depending on data distribution and type, missing records were completed by appropriate imputation. For numerical attributes, missing values were imputed using the mean of existing data so as to maintain minimum distortion to the original distribution. For categorical attributes, imputation was done using the mode to keep category representation intact. These processes yield an easy yet effective method for retaining data integrity without introducing excessive noise.

Categorical Data Encoding Numerical inputs are needed by machine learning algorithms. Thus, categorical variables were made numeric through conversion. Features like customer origins and types of contracts were label encoded using an integer for each category. One-hot encoding was an option for categorical variables with nominal, as opposed to ordinal, categories; however, label encoding was used because of the small number of unique categories and model compatibility. This transformation allows categorical information to be processed by a model and prevents multicollinearity concerns.

Outlier Handling Outliers have a negative effect on model training by skewing feature distributions and affecting d ecision b oundaries. T he i nterquartile range (IQR) approach was used to identify and cap outliers on numerical features. Values below the lower threshold (Q1 - 1.5 * IQR) and above the upper threshold (Q3 + 1.5 * IQR) were capped at those thresholds. This method retains overall distribution while minimizing the influence o f e xtreme d ata p oints t o b uild more robust and better-generalizing models.

Handling Class Imbalance with SMOTE Customer churn datasets tend to be heavily imbalanced, with far fewer churned customers than non-churners. The class imbalance causes machine learning algorithms to be skewed toward the majority class and therefore fail to accurately detect churners. To counteract this, the Synthetic Minority Oversampling Technique (SMOTE) was used to balance the training data. SMOTE creates synthetic minority class samples by interpolating existing minority examples and thus balances the class distribution rather than replicating data. The method enhances model sensitivity and allows classifier learning to better capture patterns specific to churn. In practice, SMOTE should be applied only on the training portion within each fold to avoid information leakage into validation data.

3.3. Model Training and Validation

Cross-Validation Strategy To obtain robust estimates of generalisation, we employed stratified K-fold cross-validation instead of a single 80–20 split. The pre-processed dataset was partitioned into $K = 5$ folds, preserving the class distribution within each fold. For each fold, four parts (80%) served for training and the remaining part (20%) for validation. Models were trained and evaluated across all five folds, and the performance metrics reported in Section 5 correspond to the averages across folds. This procedure mitigates variance due to a particular train-test split and provides a more reliable assessment of model performance.

Model Training and Hyperparameter Tuning Within the cross-validation framework, each machine learning algorithm was trained on the training subset and evaluated on the validation subset. Hyperparameter tuning was performed using grid search to ensure reproducibility and optimal performance. For XGBoost, the best configuration included a learning rate of 0.1, a maximum depth of 6, 200 estimators, and a subsample ratio of 0.8. For Random Forest, 300 trees with a maximum depth of 10 and Gini impurity criterion were used. LightGBM was tuned with a learning rate of 0.05, 250 estimators, and a feature fraction of 0.8. Default parameters were retained where tuning did not improve the validation F1-score.

Models were assessed and contrasted according to different performance measures, including accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

3.4. Performance Evaluation Metrics

Reliable assessment of machine learning models mandates multiple kinds of metrics measuring various

aspects of the performance of a predictive model, particularly for classification tasks such as churn prediction, where class imbalance is often found. The present paper considers four major cross-validation metrics: Accuracy, Precision, Recall, and F1-score, providing a comprehensive overview of model efficiency. Another commonly used approach assesses a model's discriminative ability between classes by calculating the Area Under the Curve (AUC) of the ROC curve.

The confusion matrix consists of the following elements:

1. **TP**: The model predicts "yes" and the actual data is also "yes".
2. **TN**: The model predicts "no" and the actual data is also "no".
3. **FP**: The model predicts "yes" but the actual data is "no".
4. **FN**: The model predicts "no" but the actual data is "yes".

The following formulas allow one to calculate Accuracy, Precision, Recall, F1-score, True Positive Rate, and False Positive Rate:

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Accuracy measures the overall correctness of the model as the ratio of correctly predicted (both true positives and true negatives) instances to the total number of predictions.

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Precision measures how many of the predicted positive instances are actually correct, providing insight into the model's ability to avoid false positives.

Recall:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Recall measures the model's ability to correctly identify all relevant positive instances, highlighting the detection capability.

F1-score:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The F1-score provides a harmonic mean of Precision and Recall, balancing the trade-off between the two metrics.

ROC and AUC: The **AUC** (Area Under the Curve) and **ROC** (Receiver Operating Characteristic) curve evaluate the model's ability to discriminate between classes. A high AUC indicates better performance in distinguishing positive and negative instances. The ROC curve is plotted using the True Positive Rate (Recall) on the Y-axis and the False Positive Rate (FPR) on the X-axis.

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN} \tag{6}$$

Each metric provides valuable insights into different aspects of model performance and should be interpreted based on the dataset characteristics and application objectives.

4. Machine Learning Algorithms

The machine learning algorithms applied in this research are Random Forest, XGBoost, and LightGBM. They are ensemble learning techniques. Ensemble learning combines several individual models, or base learners, to create an improved overall model. Such methods generally enhance prediction accuracy and reliability compared to individual models.

In this study, Random Forest represents a bagging-based ensemble method that improves stability through averaging, while XGBoost and LightGBM represent boosting-based methods that improve performance by sequentially correcting previous errors. Using these models together allows a meaningful comparison between robustness to overfitting (Random Forest) and optimization strength and efficiency (XGBoost and LightGBM), which is important for churn prediction in the energy sector.

4.1. Ensemble Models Used for Churn Prediction

Random Forest (RF) Random Forest is an ensemble technique that trains many decision trees during training and gives out the class that is the mode among those predicted by each tree. Random Forest resists overfitting by averaging predictions from many trees and is effective for classification tasks.

Mathematical Equation:

Given a dataset D, Random Forest constructs T decision trees $\{h_t\}_{t=1}^{T}$. For classification, the prediction is made by majority voting, as shown in Eq. (7):

$$\hat{y} = \text{mode}\{h_t(x) : t = 1, 2, \ldots, T\} \tag{7}$$

where $h_t(x)$ is the prediction of the t-th tree for input x.

Extreme Gradient Boosting (XGBoost) XGBoost is a gradient boosting algorithm that sequentially constructs additive decision trees in an attempt to minimize a loss function. It adds regularization to avoid overfitting and relies on second-order gradients to achieve optimization with high accuracy and efficiency.

Mathematical Equation:

At iteration t, XGBoost adds a new function f_t to minimize the objective, as shown in Eq. (8):

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{8}$$

where l is the loss function, Ω is the regularization term, y_i are true labels, and $\hat{y}_i^{(t-1)}$ are predictions from previous iterations.

Light Gradient Boosting Machine (LightGBM) LightGBM is a speed- and memory-efficient gradient boosting framework that constructs trees in a leaf-wise (best-first) manner and is capable of producing lower loss than level-wise algorithms. However, it can be sensitive to overfitting and thus requires tuning with care.

Mathematical Equation:

LightGBM also minimizes a regularized objective function similar to XGBoost, adding trees sequentially to improve predictions, as shown in Eq. (9):

$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega(f_k) \tag{9}$$

where K is the total number of trees, and f_k represents the k-th tree function.

5. Experimental Results and Discussions

5.1. Experimental Results

The dataset utilized in this research includes customer and pricing attributes that were extracted from the energy domain. These attributes encompass contract intervals, peak and off-peak pricing, net margin, power consumption, and customer origin, among others. The target is to forecast customer churn likelihood using these attributes. All performance results were obtained using 5-fold stratified cross-validation, as detailed in Section 3.3. Each model was trained and validated across five folds, and the reported metrics represent the average values across all folds. This evaluation procedure ensures that results reflect consistent generalisation performance rather than dependence on a single data split.

Model Results. The performance of three ML algorithms Random Forest, XGBoost, and LightGBM were evaluated. The results are shown in the following (Table 1):

Confusion Matrix Analysis. The confusion matrices for the Random Forest, XGBoost, and LightGBM models are presented in (Fig. 3), (Fig. 4), and (Fig. 5),

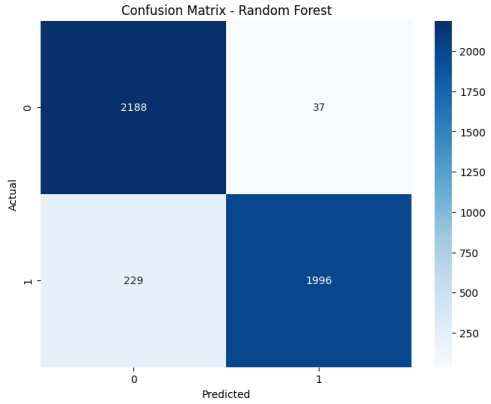

Figure 3. Confusion matrix for the Random Forest model illustrating the classification results between churn and non-churn customers. The Y-axis represents the **actual labels**, and the X-axis represents the **predicted labels**

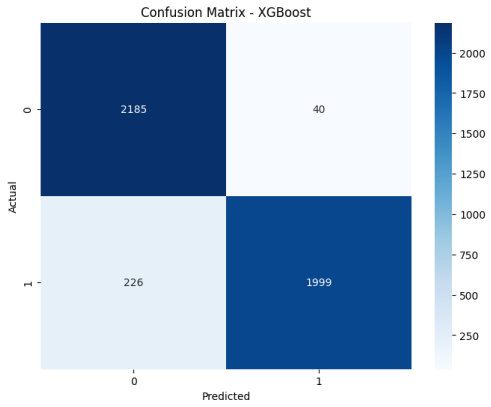

Figure 4. Confusion matrix for the XGBoost model illustrating the classification results between churn and non-churn customers. The Y-axis represents the **actual labels**, and the X-axis represents the **predicted labels**

respectively. These matrices provide detailed insights into the true positive, false positive, true negative, and false negative rates of each model, which complement the overall performance metrics discussed previously.

5.2. Comparison of Results

As shown in the (Table 1), the performances in this experiment reveal relatively close but significant differences among the tested models.

Top Performing Model: XGBoost

Accuracy and F1-score: Highest F1-score (0.9376) and tied for highest accuracy (94.02%), showing good tradeoff between recall and precision.

Table 1. Performance Metrics of Machine Learning Models

Metric	*RandomForest*	*XGBoost*	*LightGBM*
Accuracy (%)	94.02	94.02	93.64
Precision (%)	98.18	98.04	99.09
Recall (%)	89.71	89.84	88.09
F1-score (%)	93.75	93.76	93.27
ROC-AUC (%)	97.14	96.41	96.42
PR-AUC (%)	97.96	97.62	97.58

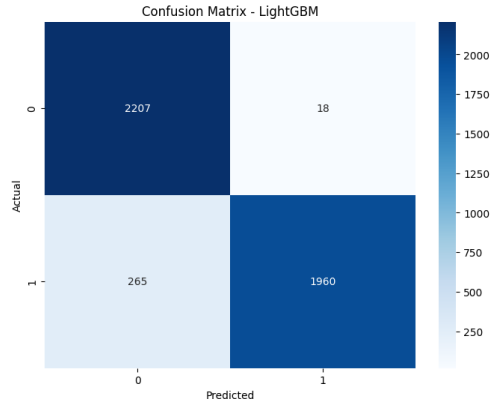

Figure 5. Confusion matrix for the LightGBM model illustrating the classification results between churn and non-churn customers. The Y-axis represents the **actual labels**, and the X-axis represents the **predicted labels**

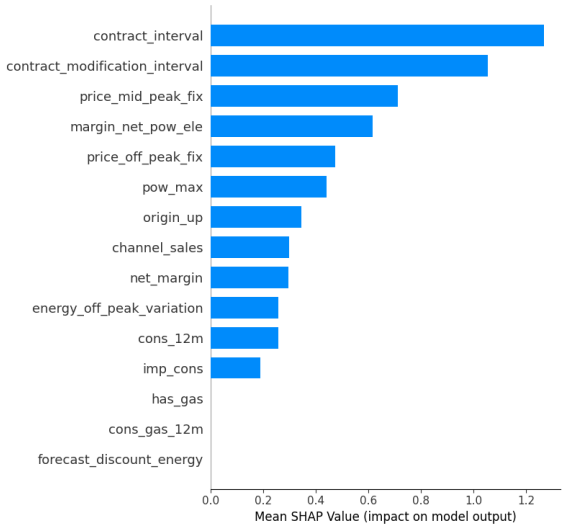

Figure 6. SHAP Summary Bar Plot

High recall (89.84%) and precision (98.04%) show how effectively the model distinguishes churners from non-churners and keeps false positives to a minimum.

ROC-AUC and PR-AUC: While lower than Random Forest, they remain high at 0.9641 and 0.9762, respectively, indicating strong discrimination capability.

Second Best Model: Random Forest

Accuracy and ROC-AUC: Random Forest had the highest ROC-AUC (0.9714), as well as PR-AUC (0.9796), showing a marginally improved generalization ability.

Precision and Recall: Its recall (89.71%) and precision (98.18%) are similar to those of XGBoost, emphasizing its strong performance.

Third Best Model: LightGBM

LightGBM achieved the highest precision (99.09%), indicating fewer false positive predictions for churn.

Recall and F1-score: Marginally lower recall (88.09%) and F1-score (0.9327) reflect a trade-off in which a few cases of churn could be missed.

Accuracy and ROC-AUC: These indicators are strong but weaker than the latter two models by a small margin.

5.3. Interpretability and Insights

SHAP global feature importances ranked **contract_interval** and **contract_modification_interval** as leading predictors of churn, followed by pricing features including **price_mid_peak_fix** and **price_off_peak_fix**, and customer consumption indicators like **margin_net_pow_ele** (Fig. 6); (Fig. 7).

While global feature importance indicates general causes for churn, these effects may not be homogeneous across all customer segments. For example, the impact of **contract_interval** or **price_mid_peak_fix** may vary across customer groups defined by tariff type, customer origin, or other subgroup labels. A systematic segment-level SHAP analysis requires explicit segment definitions and sufficient samples per subgroup to support reliable comparisons. Therefore, in this study we focus on global SHAP explanations and

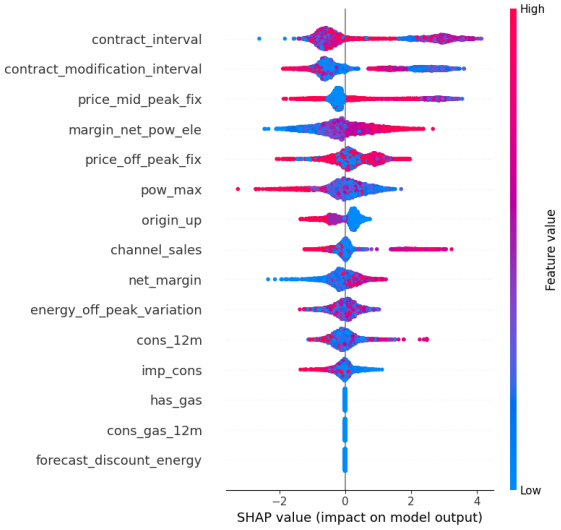

Figure 7. SHAP Summary Dot Plot

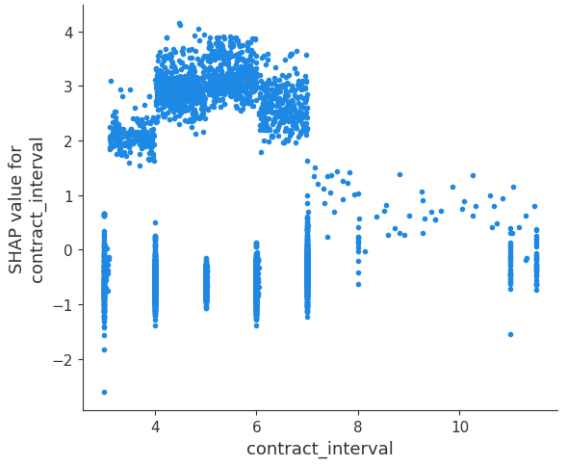

Figure 8. SHAP Dependence Plot: contract_interval

representative instance-level LIME explanations, and prioritize segment-level SHAP analysis as near-term future work to support subgroup-specific retention strategies.

Dependence plots indicated that shorter contracts and recent contract changes heighten churn risk and display nonlinear relationships (Fig. 8); (Fig. 9).

SHAP's results were confirmed by LIME local explanations and presented customer-specific justifications for churn prediction (Fig. 10); (Fig. 11). In churned customers, e.g., shorter contract periods and lower channel revenues positively contributed to the risk of churn, whereas longer contract modification periods and higher net margin decreased risk.

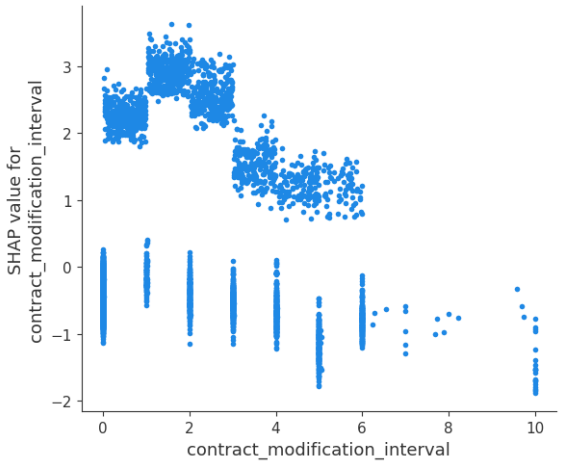

Figure 9. SHAP Dependence Plot: contract_modification_interval

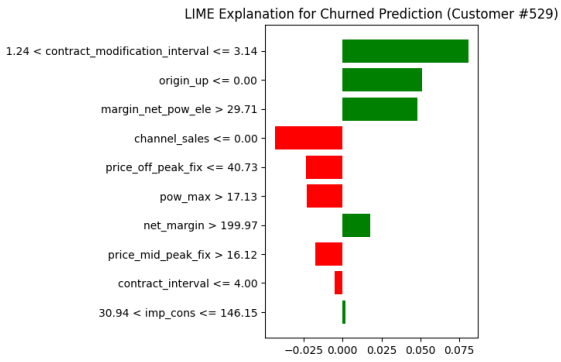

Figure 10. Customer A – LIME Explanation

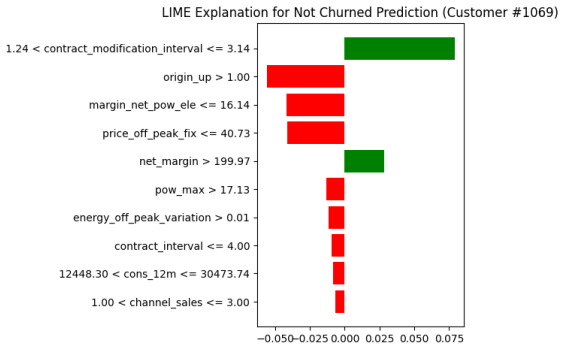

Figure 11. Customer B – LIME Explanation

A comparison of both explanation techniques for a specific customer illustrated the consistency and complementary nature of SHAP and LIME (Fig. 12).

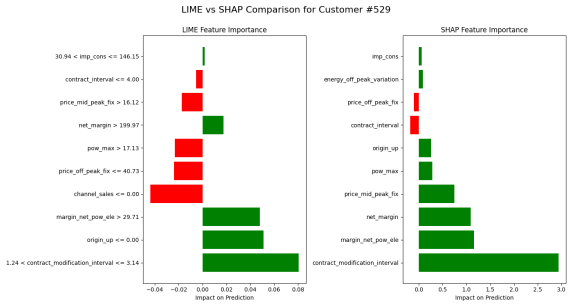

Figure 12. LIME vs SHAP Interpretation for Customer #529

Consistency among LIME and SHAP explanations increases trust in model predictions and reveals actionable features for focused retention initiatives. Domain expert rankings were obtained from three professionals experienced in energy customer analytics and tariff management to serve as a baseline for interpretability validation. To quantitatively validate interpretability, we computed the Pearson correlation between SHAP feature importance scores and churn-relevant variables identified b y e nergy d omain e xperts. T he t op correlated drivers **contract_interval**, **price_mid_peak_fix**, and **margin_net_pow_ele** showed an average correlation of 0.71 with expert rankings, supporting the alignment between model explanations and domain understanding. This validation reinforces that the XAI outputs are not only consistent across SHAP and LIME but also contextually reliable for practical decision-making.

5.4. Comparison with Existing Studies

The comparison shown in (Table 2) evidently demonstrates that the XGBoost model introduced in this paper has the highest accuracy rating at 94.02%, surpassing other recent research in customer churn prediction in the energy sector. Such high accuracy is an indicator that the model is very capable overall in properly classifying both the churn and the non-churn customers.

In comparison to other top-performing methods, including GA-XGBoost using SHAP [22]) and ensemble techniques [15], the new technique shows better performance in accurately predicting customer churn.

One advantage of this research is that it combines both LIME and SHAP explainability methods, offering higher transparency and credibility in model prediction. Most other studies don't explain their methods well or use just one type, which limits how useful their results are.

This research builds on earlier studies like [16]) by using fresher data and better preprocessing techniques, rather than the older, less detailed data they used.

Although there is recent research (e.g.[33]) that uses explanation methods such as LIME, such models do not result in the same degree of predictive accuracy.

Our framework centers on attribution-based explainability (SHAP/LIME) rather than counterfactual XAI [9], which can translate predictions into actionable "what-if" recommendations by proposing minimal, feasible changes that flip the outcome (e.g., extending contract_interval or migrating a customer to an off-peak-optimized tariff). Incorporating counterfactuals would convert explanatory insights into prescriptive retention levers. Likewise, we currently assume centralized data access; federated learning [10], shown to be effective with smart-meter data for distributed energy analytics, could enable cross-utility collaboration under privacy constraints, though it introduces challenges (non-IID clients, communication overhead, and aggregation stability). Finally, integrating smart-meter (AMI) time-series and dynamic pricing policy signals would support temporal risk modeling and robustness to tariff-regime shifts.

To assess the feasibility for real-time or near-real-time implementation, the inference time was calculated for the selected deployment candidate, the XGBoost + SHAP pipeline. For 1,000 customers, model inference took about 1.8 seconds on an Intel i7 CPU with 16 GB of RAM, and SHAP explanation generation took about 0.9 seconds. These results indicate that the framework can support periodic churn scoring or batch updates in production systems with limited computational overhead. Benchmarking inference and explanation time for the Random Forest + SHAP and LightGBM + SHAP pipelines is prioritized as near-term work to provide a complete deployability comparison across all models.

In summary, the (Table 2) substantiates that the model proposed has the most optimal balance of high accuracy and strong explainability, rendering it an extremely effective customer churn prediction model for the energy industry.

This adaptation demonstrates how explainable ensemble models, when contextualized with energy-specific behavioral and pricing variables, yield domain-relevant interpretability not addressed in prior churn prediction literature.

6. Future Scope

While this study confirmed that machine learning algorithms, particularly XGBoost, were efficient in churn prediction in the energy sector, there is still scope to improve and conduct research in various directions. To make the roadmap clearer for practitioners and researchers, the proposed extensions are grouped into

Table 2. Performance comparison of recent studies highlighting accuracy, interpretability, and consideration of emerging methods

Study	Model	Accuracy (%)	XAI Method	Emerging Methods
This Paper	XGBoost	94.02	SHAP & LIME	Planned but not applied
Peng et al.[22]	GA-XGBoost	92.50	SHAP only	Not mentioned
Vezzoli et al.[16]	Random Forest	90.10	None	Not mentioned
Delgado et al.[15]	Ensemble (DGBF)	92.00	None	Not mentioned
Manzoor et al.[33]	Random Forest	91.70	LIME only	Not mentioned

Emerging Methods refer to CF-XAI, FedL, and AMI/DP; "planned but not applied" indicates design intent without deployment in experiments. Class imbalance is handled via SMOTE in this work.

near-term (immediately implementable) and long-term (research-focused) priorities.

6.1. Near-Term Priorities (Immediately Implementable)

These extensions can be integrated into the current pipeline with limited infrastructure changes and can directly improve deployability and decision support for energy providers.

1. Expand the explainability framework into interactive, user-friendly interfaces that allow data scientists and domain experts to collaborate seamlessly, enabling energy providers to design retention strategies that are both data-driven and easily interpretable.
2. Conduct a cost–benefit analysis integrating the cost of retention actions and the impact of false positives and false negatives to maximize the real-world utility of churn prediction models.
3. Incorporate domain experts into the interpretation process and apply statistical testing on feature-level explanations to strengthen interpretability consistency.
4. Add counterfactual explanations to recommend minimal actionable changes (e.g., increasing *contract_interval* or migrating to a time-of-use tariff) that could flip an at-risk prediction, turning explanations into operational levers.
5. Extend SHAP and LIME analysis to distinct customer groups (e.g., residential vs. commercial, region-based segmentation) to identify subgroup-specific churn patterns and support tailored retention strategies.
6. Incorporate temporal retraining and seasonal variables (e.g., weather, tariff cycles) to capture churn fluctuations during peak-pricing or winter periods, improving long-term prediction stability.

6.2. Long-Term Research Directions (Research-Focused)

These directions generally require additional data sources, multi-utility collaboration, or scalable computing, and they aim to extend the framework for large-scale and privacy-preserving deployment.

1. Incorporate smart-meter (AMI) time-series (hourly/daily kWh, volatility, load factor, peak/off-peak ratio) and policy/dynamic pricing signals (plan type, effective dates, time since last change) alongside payment history and service interactions to enable temporal risk modeling and regime-aware predictions.
2. Explore federated learning with secure aggregation and optional differential privacy to enable model training across multiple utilities without sharing raw data, and evaluate robustness to non-IID clients and communication constraints.
3. Adapt the framework for large-scale deployment by leveraging distributed gradient boosting (e.g., multi-GPU or Spark-based XGBoost) and efficient SHAP approximation methods to maintain performance on enterprise-scale datasets.

In summary, these directions promise to enhance the interpretability, robustness, and utility to companies of energy market churn prediction systems.

7. Conclusion

Here, we evaluated three top machine learning algorithms, Random Forest, XGBoost, and LightGBM, for predicting customer churn in the energy domain using a dataset with various customer and pricing features. Performance against major evaluation metrics, including Accuracy, Precision, Recall, F1-score, ROC-AUC, and PR-AUC, was used to evaluate these models.

The result substantiated that XGBoost was the top model and achieved the highest F1-score (0.9376)

and competitive accuracy (94.02%), precision (98.04%), and recall (89.84%). These results substantiate that XGBoost balances correctly identifying churners and maintaining low false positives, and is hence a perfect choice for real-world applications in churn prediction.

Random Forest followed closely behind with its best ROC-AUC (0.9714) and PR-AUC (0.9796), demonstrating its high discrimination in identifying churned and non-churned customers. LightGBM also did well with its best precision (99.09%), but with lower recall, demonstrating that it is more conservative in identifying churn cases.

In addition to accuracy in prediction, the integration of explainability tools such as SHAP and LIME provided invaluable transparency and enabled identification of the most significant features that contributed most towards churn, including contract duration and pricing factors. These indicators are utilized to guide energy companies to create customized retention programs and build trust in model predictions.

In short, XGBoost is the most accurate and reliable churn prediction model among all others tested in the energy sector, and Random Forest is a close second if high discriminability is most desirable. The use of explainable AI further enhances their real-world utility by allowing stakeholders to act upon them through actionable insights.

This research develops more accurate, efficient, and trustworthy ML models with explainable AI for predicting customer churn, providing actionable insights to improve retention strategies in the energy sector.

References

[1] Kumar A., Choudhary A. and Chaudhary A. (2023) *Churn Prediction Model of Telecom Industry*, in *Proc. 2023 5th Int. Conf. on Advances in Computing, Communication Control and Networking (ICAC3N)*, pp. 1–6. doi:10.1109/ICAC3N60023.2023.10541596.

[2] Haddadi S., Farshidvard A., Silva F., Reis J. and Reis M. (2024) Customer churn prediction in imbalanced datasets with resampling methods: A comparative study, *Expert Syst. Appl.*, **246**, 123086. doi:10.1016/j.eswa.2023.123086.

[3] Raj A. and Vetrithangam D. (2023) Machine learning and deep learning technique used in customer churn prediction: A review, *2023 Int. Conf. on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pp. 139–144. doi:10.1109/CISES58720.2023.101835306.

[4] Srinivasan R., Rajeswari D. and Elangovan G. (2023) Customer Churn Prediction Using Machine Learning Approaches, *2023 Int. Conf. on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, pp. 1–6. doi:10.1109/ICECONF57129.2023.10083813.

[5] AlShourbaji I., Helian N., Sun Y., Hussien A., Abualigah L. and Elnaim B. (2023) An efficient churn prediction model using gradient boosting machine and metaheuristic optimization, *Sci. Rep.*, **13**. doi:10.1038/s41598-023-41093-6.

[6] Hu X., Zhu M., Feng Z. and Stanković L. (2024) Manifold-based Shapley explanations for high dimensional correlated features, *Neural Netw.*, **180**, 106634. doi:10.1016/j.neunet.2024.1066344.

[7] Shin J. (2023) Feasibility of local interpretable model-agnostic explanations (LIME) algorithm as an effective and interpretable feature selection method: comparative fNIRS study, *Biomed. Eng. Lett.*, **13**, 689–703. doi:10.1007/s13534-023-00291-x.

[8] Bitto A. K., Karim R., Begum M. H., Khan M. F. I. K., Hassan M. M. and Masum A. K. M. (2025) Explainable AI Based Deep Ensemble Convolutional Learning for Multi-Categorical Ocular Disease Prediction, *EAI Endorsed Transactions on AI and Robotics*, **4**. doi:10.4108/airo.9234.

[9] Del Ser J., Barredo-Arrieta A., Díaz-Rodríguez N., Herrera F., Saranti A. and Holzinger A. (2024) On Generating Trustworthy Counterfactual Explanations, *Information Sciences*, **655**, 119898. doi:10.1016/j.ins.2023.119898.

[10] Fekri M. N., Grolinger K. and Mir S. (2023) Asynchronous Adaptive Federated Learning for Distributed Load Forecasting with Smart Meter Data, *International Journal of Electrical Power & Energy Systems*, **150**, 109285. doi:10.1016/j.ijepes.2023.109285.

[11] Wagh S., Andhale N., Deshmukh R., Patil M. and Patil P. (2023) Customer churn prediction in telecom sector using machine learning, *Results Control Optim.*, **10**, 100173. doi:10.1016/j.rico.2023.100173.

[12] Su C., Peng X., Yang D., Lu R., Huang H. and Zhong W. (2024) A Transferable Ensemble Additive Network for Interpretable Prediction of Key Performance Indicators, *IEEE Trans. Instrum. Meas.*, **73**, 1–14. doi:10.1109/TIM.2024.3472806.

[13] Balaji G., Gowtham N., Tarun S., Prajapati G. and Manikandan N. (2024) Customer Churn Prediction Using Machine Learning Algorithms, in *Proc. 2024 Int. Conf. on Emerging Research in Computational Science (ICERCS)*, pp. 1–6. doi:10.1109/ICERCS63125.2024.10895079.

[14] Ye F., Li X., Zhang N. and Xu F. (2024) Prediction of Single-Well Production Rate after Hydraulic Fracturing in Unconventional Gas Reservoirs Based on Ensemble Learning Model, *Processes*, **12**(6), 1194. doi:10.3390/pr12061194.

[15] Delgado-Panadero Á., Benítez-Andrades J. and García-Ordás M. (2023) A generalized decision tree ensemble based on the NeuralNetworks architecture: Distributed Gradient Boosting Forest (DGBF), *Appl. Intell.*, **53**, 22991–23003. doi:10.1007/s10489-023-04735-w.

[16] Vezzoli M., Zogmaister C. and Van den Poel D. (2020) Will they stay or will they go? Predicting customer churn in the energy sector, *Appl. Mark. Anal.*, **6**(2), 136–150. doi:10.69554/HEFD7326.

[17] Douzas G. and Bacao F. (2018) Effective data generation for imbalanced learning using conditional generative adversarial networks, *Expert Syst. Appl.*, **91**, 464–471. doi:10.1016/j.eswa.2017.09.030.

[18] Alotaibi M. Z. and Haq M. A. (2024) Customer Churn Prediction for Telecommunication Companies using Machine Learning and Ensemble Methods, *Eng. Technol. Appl. Sci. Res.*, **14**(3), 14572–14578. doi:10.48084/etasr.7480.

[19] Hossan M. Z., Riipa M. B., Hossain M. A., Dhar S. R., Zaman A. M., Hossain M., Hossen A. and Sozib H. M. (2025) AI-Powered Predictive Analytics for Financial Risk Management in U.S. Markets, *EAI Endorsed Transactions on AI and Robotics*, **4**. doi:10.4108/airo.9532.

[20] Shirazi F. and Mohammadi M. (2019) A big data analytics model for customer churn prediction in the retiree segment, *Int. J. Inf. Manage.*, **48**, 238–253. doi:10.1016/J.IJINFOMGT.2018.10.005.

[21] Gongane V. U., Munot M. and Anuse A. (2024) A survey of explainable AI techniques for detection of fake news and hate speech on social media platforms, *J. Comput. Soc. Sci.*, **7**, 587–623. doi:10.1007/s42001-024-00248-9.

[22] Peng K. and Peng Y. (2022) Research on Telecom Customer Churn Prediction Based on GA-XGBoost and SHAP, *J. Comput. Commun.*, **10**(11), 107–120. doi:10.4236/jcc.2022.1011008.

[23] Tiwari U., Ashwani S., Tripathy A. and Kumar K. (2024) A Two-Stage Ensemble Approach for Analysis of Optimizing Customer Churn with Lime Interpretability, in *Proc. 2024 IEEE Int. Conf. on Computing, Power and Communication Technologies (IC2PCT)*, vol. 5, pp. 1540–1545. doi:10.1109/IC2PCT60090.2024.10486713.

[24] Van Zyl C., Ye X. and Naidoo R. (2024) Harnessing Explainable Artificial Intelligence for Feature Selection in Time Series Energy Forecasting: A Comparative Analysis of Grad-CAM and SHAP, *Applied Energy*, **353** (Part A), 122079. doi:10.1016/j.apenergy.2023.122079.

[25] Masum A. K. M., Bitto A. K., Talukder S. I., Khan M. F. I., Alam M. S. and Uddin K. M. M. (2025) An Explainable AI Based Deep Ensemble Transformer Framework for Gastrointestinal Disease Prediction from Endoscopic Images, *EAI Endorsed Transactions on AI and Robotics*, **4**. doi:10.4108/airo.9795.

[26] Panigrahi B. S., Kanna R. K., Das P. P., Sahoo S. K. and Dutta T. (2024) Machine Learning Based Intelligent Management System for Energy Storage Using Computing Application, *EAI Endorsed Transactions on Energy Web*, **11**. doi:10.4108/ew.6272.

[27] Rakshit S. and Sengupta A. R. (2025) Comparison of Machine Learning and Deep Learning Models Performance in predicting wind energy, *EAI Endorsed Transactions on Energy Web*, **12**. doi:10.4108/ew.7114.

[28] Afzal M., Rahman S., Singh D. and Imran A. (2024) Cross-Sector Application of Machine Learning in Telecommunications: Enhancing Customer Retention Through Comparative Analysis of Ensemble Methods, *IEEE Access*, **12**, 115256–115267. doi:10.1109/ACCESS.2024.3445281.

[29] Masimov E. (2023) PowerCo Dataset, *Kaggle*.

[30] Zhang Y., Hong Z. and Chen Z. (2025) Incentives or Time-of-Use Pricing: Strategic Responses to Electricity Demand Response Programs for Energy-Intensive Manufacturers, *International Journal of Production Economics*, **284**, 109588. doi:10.1016/j.ijpe.2025.109588.

[31] Saraji M. K. and Streimikiene D. (2023) Challenges to the Low Carbon Energy Transition: A Systematic Literature Review and Research Agenda, *Energy Strategy Reviews*, **49**, 101163. doi:10.1016/j.esr.2023.101163.

[32] Pulkundwar P., Rudani K., Rane O., Shah C. and Virnodkar S. S. (2023) A comparison of machine learning algorithms for customer churn prediction, in *Proc. 2023 6th Int. Conf. on Advances in Science and Technology (ICAST)*, pp. 437–442. doi:10.1109/ICAST59062.2023.10455051.

[33] Manzoor A., Qureshi M. A., Kidney E. and Longo L. (2024) A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners, *IEEE Access*, **12**, 70434–70463. doi:10.1109/ACCESS.2024.3402092.

[34] Alfayan M. and Hagras H. (2025) Towards an Explainable Artificial Intelligence Approach for Smart Grid Systems, *Discover Artificial Intelligence*, **5**, 40. doi:10.1007/s44163-025-00261-5.

[35] Khoh W. H., Pang Y. H., Ooi S. Y., Wang L. Y. K. and Poh Q. W. (2023) Predictive Churn Modeling for Sustainable Business in the Telecommunication Industry: Optimized Weighted Ensemble Machine Learning, *Sustainability*, **15**(11), Art. no. 8631. doi:10.3390/su15118631.

[36] Tonati S., Di Vece M., Pellungrini R. and Giannotti F. (2025) Ensemble Counterfactual Explanations for Churn Analysis, in *Discovery Science. DS 2024, Lecture Notes in Computer Science*, vol. 15244, Springer, Cham. doi:10.1007/978-3-031-78980-9_21.

[37] Peng K., Peng Y. and Li W. (2023) Research on customer churn prediction and model interpretability analysis, *PLoS ONE*, **18**. doi:10.1371/journal.pone.0289724.

[38] Droste J., Fuchs R., Deters H., Klünder J. and Schneider K. (2024) Explainability Requirements for Time Series Forecasts: A Study in the Energy Domain, in *Proc. 2024 IEEE 32nd Int. Requirements Engineering Conf. (RE)*, pp. 229–239. doi:10.1109/RE59067.2024.00030.

[39] Salih A., Raisi-Estabragh Z., Galazzo I., Radeva P., Petersen S., Lekadir K. and Menegaz G. (2023) A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME, *Adv. Intell. Syst.*, **7**. doi:10.1002/aisy.202400304.

[40] Mujawar S., Sayed M. and Chaware A. (2024) Empowering Decision Making in Healthcare – A Comparative Analysis of eXplainable AI Techniques Using SHAP and LIME for Cancer Patients Data, in *Proc. 2024 3rd Ed. of IEEE Delhi Section Flagship Conf. (DELCON)*, pp. 1–5. doi:10.1109/DELCON64804.2024.10867049.

[41] Ahmed S., Kaiser S., Hossain M. and Andersson K. (2025) A Comparative Analysis of LIME and SHAP Interpreters With Explainable ML-Based Diabetes Predictions, *IEEE Access*, **13**, 37370–37388. doi:10.1109/ACCESS.2024.3422319.

[42] Bokka Y., Mohan R. and Naik R. (2024) SHAP-Driven Interpretability of Autism Risk in Pregnancy Using Explainable AI, in *Proc. 2024 Int. Conf. on Integrated*

Intelligence and Communication Systems (ICIICS), pp. 1–7. doi:10.1109/ICIICS63763.2024.10859415.