IAJ
Indonesian Actuarial Journal
Persatuan Aktuaris Indonesia (PAI)

Vol. 01, No. 02, December (2025)
e-ISSN 3110-6463
https://10.65689/iajvol01no2pp137-151

Precision-Oriented Churn Prediction with a Fine-Tuned Meta-Learner Stack Model and SHAP: A Case Study on IBM Telco

Antoni T.G [1], Hakim A.H [1*] , Hendartriany R.N [1] , Vyorra A.V.E [1], Septyanto F [1]

[1] *School of Data Science, Mathematics, and Informatics, IPB University, Bogor, Indonesia*

haidarhakim@apps.ipb.ac.id

Abstract

Customer churn prediction is essential in the telecommunications industry, where maintaining existing customers is significantly more cost-effective than acquiring new ones. This study introduces a precision-oriented stacked ensemble model to predict churn using the IBM Telco Customer Churn dataset. Emphasis is placed on maximizing precision to reduce false positives, thereby minimizing unnecessary and costly intervention efforts. The proposed architecture employs LightGBM, CatBoost, and Logistic Regression as base learners, with a fine-tuned ElasticNet serving as the meta-learner. Evaluation results show that the stacking model achieves strong overall performance, attaining an AUC of 0.917 and the highest precision among all models tested. To ensure interpretability, SHapley Additive exPlanations (SHAP) are applied to identify key drivers of churn such as number of referrals, contract type, monthly charges, and tenure. These findings demonstrate that a precision-focused approach can balance business efficiency and predictive power, offering a robust framework for proactive and cost-sensitive churn management.

Article History:

Received: 1 May 2025

First Revised: 2 July 2025

Accepted: 15 November 2025

Published: 31 December 2025

Keywords:

Churn, Model, Stacking, SHAP, Analysis.

1. INTRODUCTION

Customer churn remains a critical challenge in the telecommunications industry, as it significantly impacts revenue stability and long-term competitiveness. Retaining existing customers is far more cost-effective than acquiring new ones, with acquisition expenses estimated to be up to five times higher than retention costs [1]. Globally, churn rates in telecom services vary between 20% and 40% annually, influenced by regional market conditions and service offerings [2]. Consequently, early and accurate identification of at-risk customers has become essential for sustaining profitability.

One promising solution is churn prediction using machine learning techniques. Traditional models such as Random Forest, Gradient Boosting, and Logistic Regression have achieved high classification performance, with Random Forest models often exceeding 85% accuracy and an F1-score above 0.80 [3]. However, single models can suffer from limited generalization on unseen data. As a result, ensemble methods have gained traction for their ability to integrate multiple models and improve prediction accuracy and stability [4], [5].

Stacked ensemble models combine diverse base learners such as decision trees, gradient boosters, and neural networks with a meta-learner that synthesizes their outputs. Research has shown that stacking can improve churn prediction performance by 3–5% over individual models, especially in imbalanced datasets [6], [7]. For instance, Shaikhsurab and Magadum [6] implemented a stacking model incorporating XGBoost, LightGBM, and neural networks, achieving a 99.28% accuracy on telecom churn datasets. Similarly, Awang et al. [7] demonstrated that stacking outperformed single classifiers, bagging, and boosting techniques across several evaluation metrics.

Although these studies validate the effectiveness of stacked models, most focus on optimizing general metrics such as accuracy or AUC. In practice, however, false positives in churn prediction may result in unnecessary intervention efforts and operational costs. This business context underscores the need for a precision-oriented approach that emphasizes minimizing false positives. Despite its relevance, only a few studies have explicitly targeted precision as the primary objective, especially in stacking-based frameworks [8].

In addition, there is limited research that combines diverse base learners such as LightGBM, CatBoost, and Logistic Regression with a regularized meta-learner like ElasticNet, while also incorporating model interpretability techniques such as SHapley Additive exPlanations (SHAP). Previous works often focus either on performance or interpretability, but rarely aim to balance both dimensions simultaneously [9].

To address these gaps, this study proposes a precision-oriented stacking ensemble model for churn prediction. The model integrates LightGBM, CatBoost, and Logistic Regression as base learners, with ElasticNet as the meta-learner. Although the model was initially tuned using AUC to ensure robust classification capability, evaluation results highlight its superior precision, making it particularly suitable for cost-sensitive scenarios. Furthermore, SHAP is employed to provide interpretability, allowing business stakeholders to identify and understand the key drivers of churn.

2. METHODS

2.1 Data Source and Preprocessing

In this study, the IBM Telco Customer Churn dataset is used, which was obtained through an open-source platform, Kaggle. The dataset contains a total of 7,043 customer entries, each consisting of 50 features representing multidimensional information about customer characteristics and churn status. The dataset was selected due to its relevance and completeness in capturing historical customer data in the telecommunications sector, including demographic variables, types of services used, and customer churn status. Preliminary analysis reveals that the dataset has an imbalanced class distribution, with 26.54% of customers having churned and 73.46% having remained, as visualized in **Figure 1.**

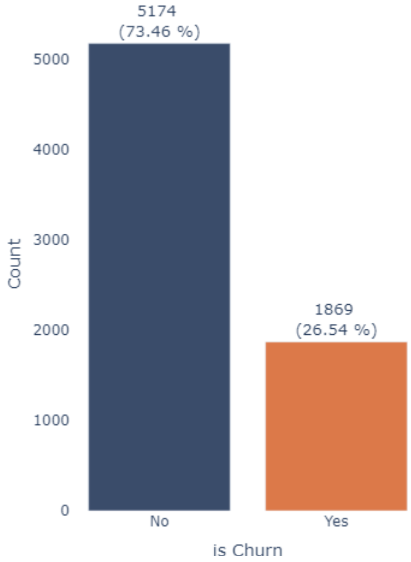

Figure 1. Distribution diagram of customer churn

This study utilizes the IBM Telco Customer Churn dataset, obtained from the open-source platform Kaggle. The dataset contains 7,043 customer records with 50 initial features, comprising 19 numerical and 31 categorical variables. These features include demographic data, customer behavior, billing history, and service subscriptions.

Prior to modeling, an extensive preprocessing phase was conducted to ensure data quality and compatibility with machine learning algorithms. As illustrated in **Figure 2.**, the preprocessing steps included:

- Feature elimination: Irrelevant or redundant variables such as 'Customer ID', 'Country', 'State', 'Latitude', 'Longitude', and high-missing-value fields like 'Churn Reason' and 'Offer' were removed.
- Handling missing values: Missing categorical values were either removed or filled with default labels (e.g., “Unknown”), depending on distribution and frequency.
- Feature transformation: Variables were transformed using different encoding strategies:
    - StandardScaler for normalizing continuous numerical features such as 'Tenure in Months', 'Monthly Charge', and 'Total Revenue'.
    - OrdinalEncoder for ordered categorical features (e.g., 'Contract', 'Senior Citizen', 'Multiple Lines').
    - OneHotEncoder for nominal categorical variables such as 'Contract', 'Payment Method', and 'Offer'.
    - CountEncoder for high-cardinality categorical features like 'City' and 'Zip Code'.

After preprocessing, the resulting dataset consisted of 47 engineered features, as visualized in **Figure 2.**, which shows the transformation from the raw (Before) to final processed dataset (After).

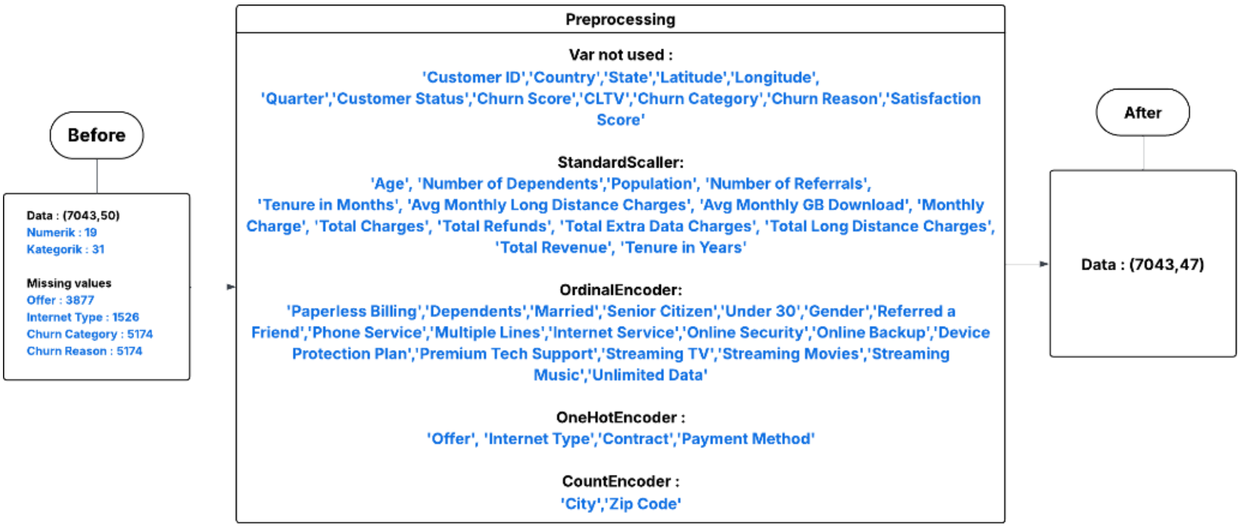
Before
Data : (7043,50)
Numerik : 19
Kategorik : 31
Missing values
Offer : 3877
Internet Type : 1526
Churn Category : 5174
Churn Reason : 5174
Preprocessing
Var not used :
'Customer ID','Country','State','Latitude','Longitude', 'Quarter','Customer Status','Churn Score','CLTV','Churn Category','Churn Reason','Satisfaction Score'
StandardScaller:
'Age', 'Number of Dependents','Population', 'Number of Referrals', 'Tenure in Months', 'Avg Monthly Long Distance Charges', 'Avg Monthly GB Download', 'Monthly Charge', 'Total Charges', 'Total Refunds', 'Total Extra Data Charges', 'Total Long Distance Charges', 'Total Revenue', 'Tenure in Years'
OrdinalEncoder:
'Paperless Billing','Dependents','Married','Senior Citizen','Under 30','Gender','Referred a Friend','Phone Service','Multiple Lines','Internet Service','Online Security','Online Backup','Device Protection Plan','Premium Tech Support','Streaming TV','Streaming Movies','Streaming Music','Unlimited Data'
OneHotEncoder :
'Offer', 'Internet Type','Contract','Payment Method'
CountEncoder :
'City','Zip Code'
After
Data : (7043,47)

Figure 2. Data preprocessing pipeline

2.2 Overall Process of Customer Churn Analysis Model Construction

This study employs a machine learning pipeline to predict customer churn, structured in several stages as illustrated in **Figure 3**. The process begins with data cleaning, imputation, and feature transformation through appropriate encoding strategies. Following preprocessing, the dataset is split into training and test sets, and further evaluated using repeated K-Fold cross-validation to ensure robust model assessment.

Several classification models were trained, including Decision Tree, K-Nearest Neighbors, Logistic Regression, Random Forest, XGBoost, Gradient Boosting, Extra Trees, LightGBM, and CatBoost. Based on their ROC AUC performance, three top-performing models were selected as base learners for a stacking ensemble.

These base learners produced out-of-fold predictions which were used as input features for the meta-learner, implemented as a Logistic Regression model. To ensure optimal performance and generalization, the meta-learner was tuned using Optuna, which explored a range of regularization settings:

- Various solver–penalty combinations, including:
  ('liblinear', 'l1'), ('liblinear', 'l2'), ('saga', 'l1'), ('saga', 'l2'), and ('saga', 'elasticnet')
- Regularization strength CCC, searched on a log-uniform scale between 10−310^{-3}10−3 and 10210^2102
- Number of iterations (max_iter) ranging from 100 to 1000
- For ElasticNet, the l1_ratio was also optimized in the range [0, 1]

All configurations were evaluated using ROC AUC as the primary optimization metric. This process ensured a fair and systematic selection of the meta-learner configuration for the final ensemble model. Model interpretability was performed using SHapley Additive exPlanations (SHAP), allowing both global and local understanding of feature contributions to churn predictions.

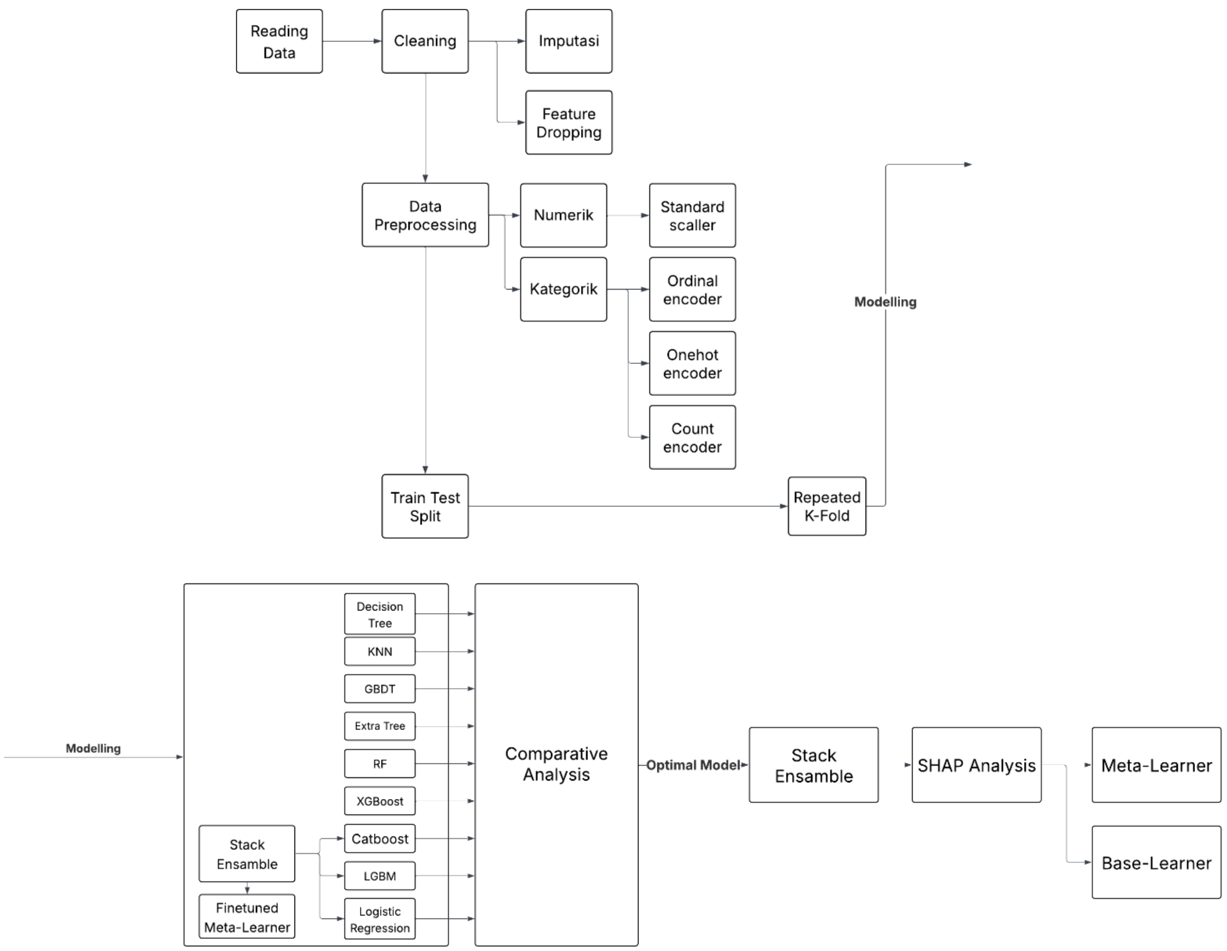

Figure 3. General flow chart of IBM telco customer churn analysis model

2.3 Machine Learning Algorithms

Four main algorithms were utilized to build the churn prediction model, including three base learners and one fine-tuned meta-learner. Our approach maximalize the potential of the model also increasing the accuracy of the overall method being used. The description of each model and its loss function is provided below:

LightGBM: Light Gradient Boosting Machine (LightGBM) is a tree-based boosting algorithm that follows a leaf-wise growth strategy and is designed for high efficiency with large datasets. It supports various data types, including categorical features, and is suitable for both classification and regression tasks. LightGBM employs advanced techniques such as early stopping, regularization, and histogram-based binning to speed up training and reduce overfitting. A 5-fold cross-validation reduces bias and ensures the model performs well with any subset of testing data [10]. LightGBM optimizes a combined objective function of loss and regularization:

$$L(\emptyset) = \sum_{i=1}^{n} l(y_i, \hat{y}_1) + \sum_{k=1}^{K} \Omega(f_k) \qquad (1)$$

Where

$$\Omega(f_k) = \gamma T \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \qquad (2)$$

with T as the number of leaves and w_j the output at leaf j. The loss function typically used is binary cross-entropy with L2 regulariza tion on the leaf outputs.

CatBoost: a boosting algorithm specifically designed to handle categorical fea tures efficiently. Unlike LightGBM, CatBoost employs symmetric (oblivious) trees, where all nodes at a given depth use the same splitting criterion. This results in balanced tree structures and more efficient computation. Furthermore, CatBoost introduces ordered boosting, a method that uses a permutation-driven technique to avoid target leakage and reduce overfitting, especially on small datasets. CatBoost is known for its ability to handle categorical features effectively without requiring manual encoding or preprocessing. It also uses various techniques to prevent overfitting, such as random permutations of feature values and gradient-based sampling of the training data [11]. Its objective function for binary classification is:

$$L(\emptyset) = -\sum_{i=1}^{n}[y_i \log(\sigma(\hat{y}_i)) + (1 - y_i)\log(1 - \sigma(\hat{y}_i))] \tag{3}$$

Where

$$\sigma(\hat{y}_i) = \frac{1}{1+e^{-\hat{y}_i}}.$$

CatBoost relies on implicit regularization through model structure and boost ing strategy without explicit penalty terms.

Logistic Regression: In machine learning, Logistic Regression is a linear classification model that models the probability of a class using the sigmoid function. The loss function is binary cross-entropy:

$$L(\emptyset) = -\frac{1}{n}\sum_{i=1}^{n}[yi \log(y\imath) + (1 - yi)\log(1 - \hat{y}\imath)] \tag{4}$$

ElasticNet (Finetuned Meta Learner): ElasticNet is a variation of Logistic Regression that incorporates a combination of L1 (Lasso) and L2 (Ridge) regularization. The ElasticNet model optimizes regression coefficients by selecting relevant variables through Lasso (L1) and shrinking the impact of insignificant variables using Ridge (L2), thereby enhancing the model's ability to predict customer churn more accurately [12]. The loss function is:

$$L(\emptyset) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(y_i) + (1 - y_i)\log(1 - y_i) + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2] \tag{5}$$

The ElasticNet penalty was found to be optimal through hyperparameter tuning using Optuna, resulting in better generalization for the meta-learner.

2.5 Interpretability of the Integrated SHAP Model

SHAP (Shapley Additive Explanations), a framework proposed by Lundberg, is a significant interpretability method derived from cooperative game theory. For each prediction sample, the model produces a predicted value, and SHAP calculates Shapley values for each feature, reflecting their individual contribution to the overall prediction. SHAP interprets a model's prediction as the sum of Shapley values for each input feature, explaining both the magnitude and direction of influence that each feature has on the final prediction.

Let x_i be the i-th sample, where the j-th feature of x_i is denoted by $x_{i,j}$. Let y_i represent the model's predicted value for this sample, and let y_{base} denote the average value of the target variable across all samples (i.e., the SHAP baseline value). The SHAP value associated with x_{ij} is denoted by $f(x_{ij})$. Then, the SHAP value is defined as:

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \cdots + f(x_{ik}) \tag{6}$$

3. Experimental Analysis

3.1 Prepocessing

This study utilized experimental data comprising 7,043 customer samples, each initially containing 50 attributes representing churn-related characteristics, final churn status (lost/not lost), and basic customer information. Following data cleaning procedures including handling of missing values, elimination of irrelevant features, and feature engineering a total of 38 variables were retained for modeling. During data preprocessing, continuous features were standardized using Standard Scaler (mean removal and variance scaling), while discrete features were encoded based on their nature: One Hot Encoder was applied to nominal attributes, and Ordinal Encoder to hierarchical categorical features. All transformations were integrated into a single pipeline through Column Transformer, resulting in a final feature set comprising 46 variables for modeling purposes.

3.2 Evaluation Index

This study applies a combination of five evaluation metrics, each serving a distinct purpose to ensure comprehensive and business-relevant assessment of model performance: Accuracy, Precision, Recall, F1-score, and AUC. While accuracy provides a baseline understanding of overall correctness, it is inadequate alone in imbalanced classification contexts like churn. Precision is emphasized given the business imperative to minimize false positives, which represent high cost if there are any misclassifications. F1-score balances precision and recall, while ROC AUC captures the discriminatory capacity of classifiers regardless of decision thresholds. This multi-metric framework ensures a holistic assessment of each model's utility. By incorporating all five metrics, this study not only captures the statistical performance of the models but also ensures alignment with **business objectives**, especially the focus on **high precision**, which is emphasized throughout this work.

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

$$Recall = \frac{TP}{TP + FN}$$

$$F1\,Score = \frac{2 \times precision\ \times recall}{precision + recall}$$

3.3 Experimental Results and Analysis

In this section, we selected nine classification models, XGBoost, CatBoost, LightGBM, Random Forest, Logistic Regression, KNN, Gradient Boosting, Decision Tree, and Extra Trees for comparison, and selected

Accuracy, AUC, Recall, Precision, F1-score as the evaluation indexes of the models, and the ROC comparison results of each classification model are shown in **Figure 4.,**

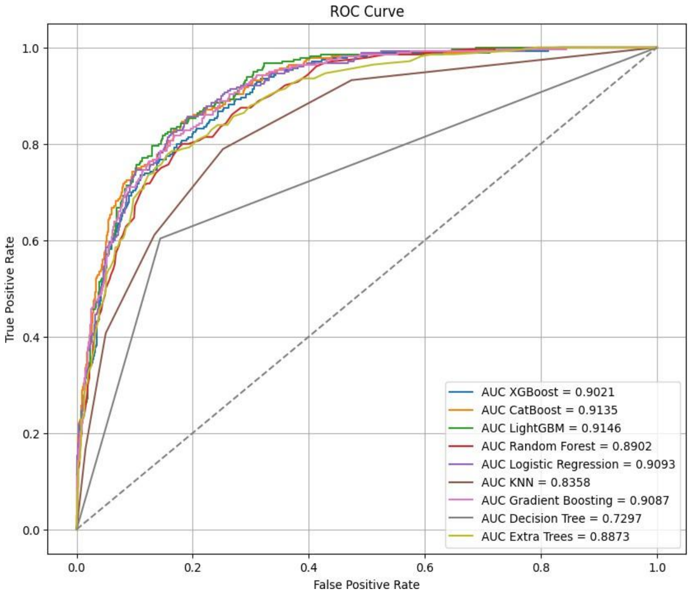

Figure 4. Comparison results of each classification model.

To enhance predictive performance, we selected three of the top-performing models: LightGBM, CatBoost, and Logistic Regression as base learners. These models independently generate **out-of-fold predictions**, where each model is trained on a subset of the training data and then used to predict the validation fold it has not seen. This process is repeated across all folds, ensuring that each prediction is made on unseen data. The resulting out-of-fold predictions are then aggregated and used as input features for the meta-learning algorithm.. As the meta-learner, we employed ElasticNet due to its ability to handle multicollinearity and perform automatic feature selection through L1 and L2 regularization. This stacking ensemble approach enables the model to capture diverse patterns learned by the base learners while reducing overfitting. The complete architecture of the stacking process is illustrated in **Figure 5.,** The out-of-fold (OOF) prediction strategy was pivotal in training the meta-learner. By using OOF outputs from the base models during cross validation, the meta-learner (ElasticNet) was exposed to unbiased base-level predictions, reducing data leakage and preserving the ensemble's generalization power.

In terms of overall predictive performance in **Figure 6.,** the models demonstrated consistently strong discrimination capabilities, with most achieving AUC values exceeding 0.90, a benchmark indicative of excellent model quality. Notably, the Stacking model outperformed all individual classifiers, achieving the highest AUC score of **0.9171**, closely followed by LightGBM (**0.9146**) and CatBoost (**0.9135**). The final approximate solution is solver = saga, penalty = elasticnet, C = 0.00412, max_iter = 388, l1_ratio = 0.665226.

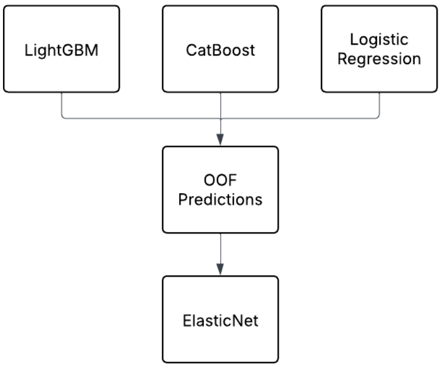

Figure 5. Final structure of stack model

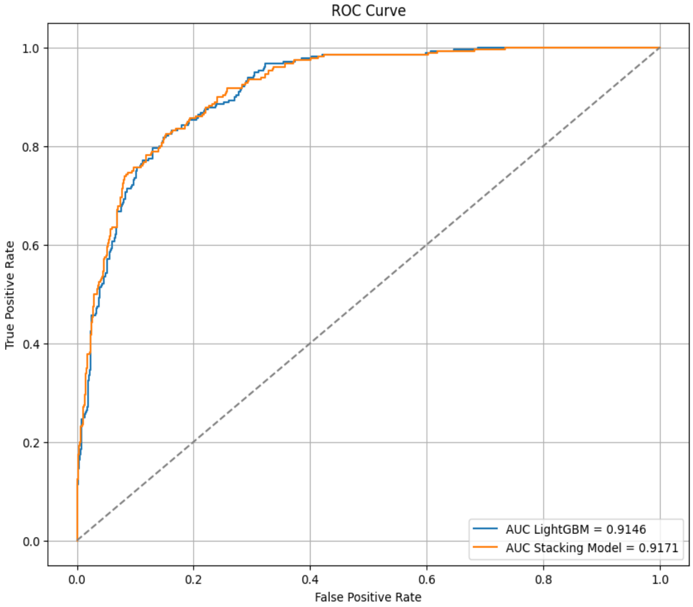

Figure 6. Comparison results of LightGBM and stacking model

Further supporting the model's effectiveness, the confusion matrix analysis in **Figure 7.,** revealed that the Stacking model attained a True Positive rate of 13.72% and maintained a low False Positive rate of just 2.65%. This optimal trade-off between sensitivity and specificity is crucial in churn-sensitive industries, where the costs of misclassification can have direct financial consequences.

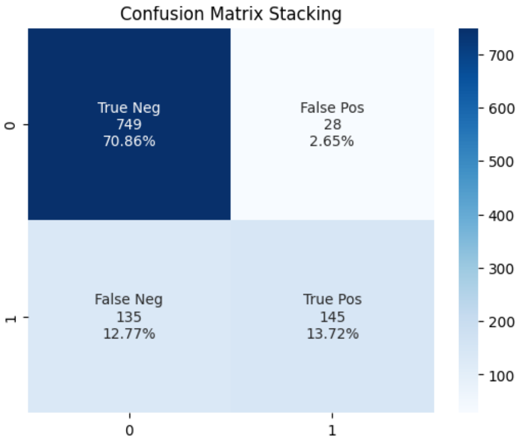

Figure 7. Confusion matrix stacking

The quantitative results in **Table 1.,** affirm our precision-oriented strategy, with the Stacking model achieving a superior Precision of 0.8382. This represents a significant 6.64% improvement over the next-best model, CatBoost (0.7860), which is critical from a business standpoint as it ensures retention efforts are cost-effectively targeted at genuine churn risks, thereby minimizing wasted resources on false positives. While other classifiers like LightGBM and CatBoost registered marginally higher accuracy or F1-scores, these came at the direct expense of lower precision. The Stacking model's performance, further validated by its class-leading AUC of 0.9171, confirms its exceptional ability to accurately identify at-risk customers, making it the most practically valuable solution for our objective.

Table 1. Model comparison value.

Model	Precision	Recall	F1-Score	Accuracy	AUC
Stacking Model	**0.8382**	0.5179	0.6402	0.8458	**0.9171**
LightGBM	0.7549	**0.6929**	0.7225	0.8590	0.9146
CatBoost	0.7860	0.6821	**0.7304**	**0.8666**	0.9135
Logistic Regression	0.7393	0.6786	0.7076	0.8515	0.9093
Gradient Boosting	0.7676	0.6607	0.7102	0.8571	0.9087
XGBoost	0.7421	0.6679	0.7030	0.8505	0.9021
Random Forest	0.7807	0.6357	0.7008	0.8562	0.8909
Extra Trees	0.7434	0.6000	0.6640	0.8392	0.8904
KNN	0.6218	0.6107	0.6162	0.7985	0.8358
Decision Tree	0.6051	0.5964	0.6007	0.7900	0.7281

Collectively, the integration of advanced ensemble modeling with detailed interpretability through SHAP analysis yields a robust, operationally viable churn prediction system. These findings provide a comprehensive and actionable framework for businesses aiming to proactively manage customer retention, demonstrating that predictive excellence and interpretability can be achieved simultaneously through a thoughtful combination of modern machine learning techniques.

3.4 Explanatory Analysis of the Model Based on SHAP

To further enhance the interpretability of our churn prediction models, this study employed SHapley Additive exPlanations (SHAP) analysis, which provides a precise measure of each feature's contribution to the model's output. SHAP values offer insight not only into the direction (positive or negative impact) of a feature on the likelihood of churn but also into the magnitude of its influence, thus serving as a transparent tool for model explanation.

The SHAP analysis was conducted individually for the three selected base learners; LightGBM, CatBoost, and Logistic Regression for the final stacking model. In each SHAP summary plot presented, the horizontal axis represents the SHAP value, indicating the extent and direction of impact, while each dot on the vertical axis corresponds to an individual customer record. The color gradient from blue to red denotes the relative magnitude of feature values, with blue indicating low feature values and red representing high feature values.

3.4.1 SHAP Summary Plot for LightGBM

The SHAP summary plot for LightGBM, as illustrated in Figure 7**.**, clearly identifies Number of Referrals as the most influential feature in reducing customer churn. Customers with higher referral counts exhibited SHAP values skewing heavily to the left, with maximum negative impacts reaching approximately **-4.0** on the SHAP scale. This strong negative contribution indicates that customers who referred others were substantially less likely to churn, suggesting social loyalty effects. Meanwhile, the **Contract_Two Year** feature exhibited consistent SHAP values between **-2.5** and **-3.0**, reinforcing the notion that customers bound by two-year contracts possess a significantly lower propensity to churn compared to others. Conversely, the **Monthly Charge** feature demonstrated a SHAP value distribution heavily leaning toward positive contributions, peaking around **+2.5**, implying that higher monthly costs notably elevate the probability of churn.

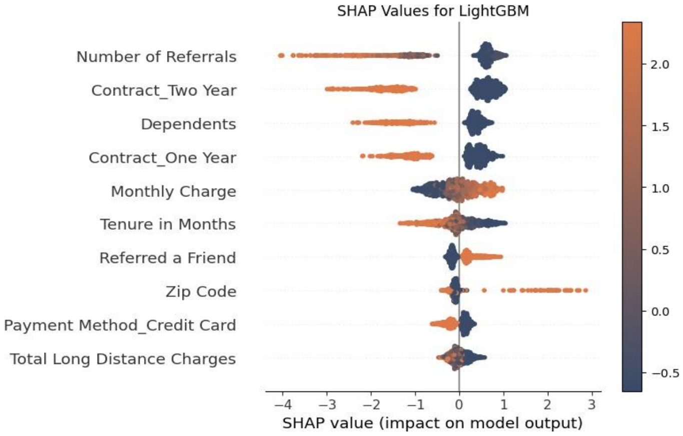

Figure 8. SHAP Summary plot for LightGBM

Additionally, Tenure in Months showed a complex pattern; lower tenures (depicted in blue) corresponded to positive SHAP values, highlighting that newer customers were more inclined to churn, while longer tenures (in red) contributed negatively to churn likelihood. Through this LightGBM-based SHAP interpretation, it becomes evident that financial commitment (via contract terms) and social factors (through referrals) are critical in customer retention strategies.

3.4.2 SHAP Summary Plot for CatBoost

In the SHAP summary plot derived from CatBoost, shown in Figure 8**.**, a broadly similar pattern emerges, albeit with some nuanced distinctions. The Number of Referrals again surfaced as the most influential feature, with SHAP values reaching negative extremes near **-2.5**. The **Contract_Two Year** feature mirrored the same protective impact against churn, with SHAP values clustering around **-2.0**, although slightly less pronounced compared to LightGBM. Interestingly, **Monthly Charge** maintained its role as a risk amplifier, but in CatBoost, the maximum positive SHAP values were slightly subdued, reaching approximately +2.0. Moreover, CatBoost revealed Number of Dependents as a moderately impactful feature, where in customers with more dependents (higher feature values marked in red) tended to show negative SHAP values, suggesting they were less likely to churn, potentially due to bundled service incentives or family plan benefits.

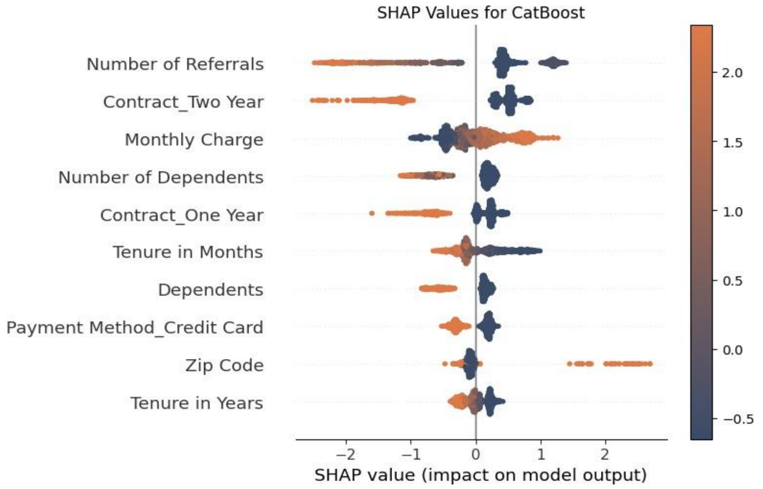

Figure 9. SHAP Summary plot for CatBoost

In the SHAP summary plot derived from CatBoost, shown in **Figure 8**, a broadly similar pattern emerges compared to LightGBM, albeit with nuanced distinctions. **Number of Referrals** remained the strongest predictor, showing high negative SHAP values near –2.5, indicating a reduced churn likelihood for socially connected users. **Contract_Two Year** continued to act as a strong retention indicator, while **Monthly Charge** again contributed positively to churn risk.

What sets CatBoost apart is its elevated attribution to features associated with personal and household context. For example, the model assigned moderate SHAP importance to **Number of Dependents**, with higher values (i.e., customers with more dependents) showing consistent negative contributions to churn. This suggests that customers with family responsibilities may value bundled or multi-line plans, aligning their churn behavior with underlying socioeconomic factors. Unlike LightGBM, which focused more on service contract and pricing-related variables, CatBoost provided stronger signals from features such as **Dependents** and **Tenure in Years**, which indirectly reflect customer life stage, family structure, and long-term economic engagement. These patterns indicate that CatBoost is more sensitive to **socioeconomic signals**, possibly due to its superior handling of categorical and ordinal variables without aggressive transformation or binning

3.4.3 SHAP Summary plot for Logistic Regression

Turning to Logistic Regression, depicted in Figure 9**.**, the SHAP distribution presents a distinctly linear interpretive pattern, aligning with the model's inherently interpretable nature. The feature Tenure in Months emerged as the most decisive in reducing churn risk, with SHAP values extending down to nearly **-4.5**, which marks the largest negative SHAP contribution among all models analyzed. This finding emphasizes that the longer a customer remains subscribed, the less likely they are to churn, a finding both intuitively sound and empirically verified. In contrast, Monthly Charge consistently displayed positive SHAP values, peaking around **+2.5**, confirming that customers facing higher charges were more susceptible to churn decisions.

Secondary features like Contract_Two Year and Referred a Friend continued to support the trend of reducing churn, albeit with more moderate SHAP values ranging between **-1.5** to **-2.0**.

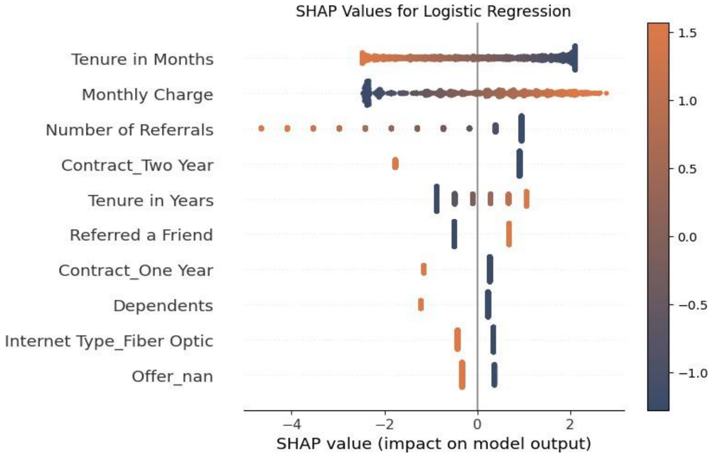

Figure 10. SHAP Summary plot for Logistic Regression

These findings highlight Logistic Regression's strength in isolating individual linear relationships, reaffirming the essential importance of service tenure and pricing in managing customer loyalty.

3.4.4 SHAP Summary plot for Stacking Model

Finally, the SHAP analysis of the Stacking Model, presented in Figure 10**.**, illustrates how the combined predictive strength of multiple learners synergizes feature contributions. Within this ensemble model, LightGBM emerged as the dominant contributor to final predictions, with SHAP values extending up to approximately **+1.5**. CatBoost contributed substantively as well, albeit slightly less, reaching maximum SHAP impacts around **+1.2**, while Logistic Regression played a supportive role, peaking at around **+0.4**. The balanced integration of these models within stacking allowed for a nuanced capture of complex nonlinear patterns from boosting algorithms while preserving the straightforward interpretability of linear models.

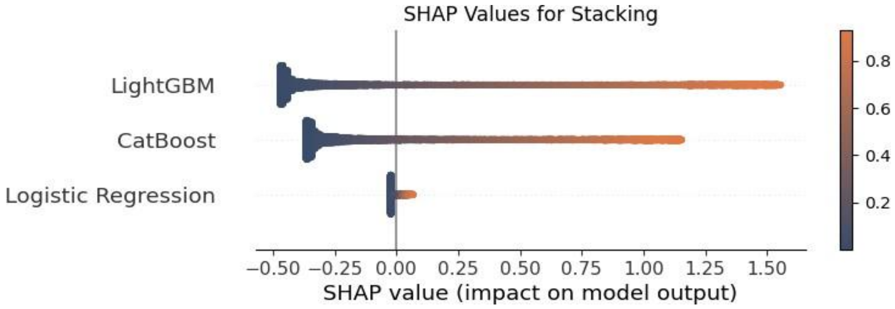

Figure 11. SHAP Summary plot for Stacking Model

This hierarchy of model contributions within the stacking framework validates the rationale behind model selection, illustrating that the blending of diverse learner perspectives produces superior and more stable predictive outcomes.

4. CONCLUSIONS

This study successfully developed a precision-oriented churn prediction framework, demonstrating that a fine-tuned stacking ensemble model integrated with SHAP explainability offers superior performance. Our primary contribution is proving that prioritizing the precision metric effectively minimizes costly false positives, directly addressing a key business challenge in customer retention. The framework excels by not only achieving high predictive accuracy but also delivering transparent, actionable insights into the primary drivers of churn, such as contract type and tenure. This provides a robust and interpretable blueprint for data-driven retention strategies that are both effective and cost-efficient.

Nevertheless, the framework's validation on the IBM Telco dataset introduces limitations in generalizability, and future work should prioritize testing on larger, real-world datasets from diverse industries. Key methodological enhancements should include exploring alternative ensemble architectures and, most importantly, incorporating cost-based evaluation metrics, such as return on intervention (ROI), to more closely align predictive performance with financial outcomes. Finally, complementing SHAP with other interpretability tools like LIME could offer a more holistic understanding of the model's behavior, strengthening stakeholder trust and its practical application in business decision-making.

5. REFERENCES

[1] K. Saleh, “Customer Acquisition Vs Retention Costs: Statistics & Trends You Should Know,” https://www.invespcro.com/blog/customer-acquisition-retention/.

[2] J. Arbanas and D. Littmann, “2023 telecom industry outlook,” May 2023. Accessed: Apr. 30, 2025.https://www.deloitte.com/global/en/Industries/tmt/perspectives/telecommunications-industry-outlook.html

[3] V. Chang, K. Hall, Q. A. Xu, F. O. Amao, M. A. Ganatra, and V. Benson, “Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models,” *Algorithms*, vol. 17, no. 6, Jun. 2024, doi: 10.3390/a17060231.

[4] J. Kunnen, M. Duchateau, Z. Van Veldhoven, and J. Vanthienen, “Benchmarking Stacking Against Other Heterogeneous Ensembles in Telecom Churn Prediction,” in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, Dec. 2020, pp. 1234–1240. doi: 10.1109/SSCI47803.2020.9308188.

[5] Y. Liu, J. Fan, J. Zhang, X. Yin, and Z. Song, “Research on telecom customer churn prediction based on ensemble learning,” *J Intell Inf Syst*, vol. 60, no. 3, pp. 759–775, Jun. 2023, doi: 10.1007/s10844-022-00739-z.

[6] M. A. Shaikhsurab and P. Magadum, “Enhancing Customer Churn Prediction in Telecommunications: An Adaptive Ensemble Learning Approach.”

[7] M. K. Awang, M. Makhtar, N. Udin, and N. Farraliza Mansor, “Improving Customer Churn Classification with Ensemble Stacking Method.” [Online]. Available: www.ijacsa.thesai.org

[8] J. Li, “Customer Churn Prediction using Machine Learning: A Case Study of E-commerce Data,” 2024.

[9] Rofik, J. Unjung, and B. Prasetiyo, “Enhancing costumer churn prediction with stacking ensemble and stratified k-fold,” *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 1, pp. 398–408, Feb. 2025, doi: 10.11591/eei.v14i1.8112.

[10] A. Odeh, Q. A. Al-Haija, A. Aref, and A. A. Taleb, “Comparative Study of CatBoost, XGBoost, and LightGBM for Enhanced URL Phishing Detection: A Performance Assessment,” *Journal of Internet Services and Information Security*, vol. 13, no. 4, pp. 1–11, Nov. 2023, doi: 10.58346/JISIS.2023.I4.001.

[11] S. Leonelli and N. Tempini, “Data Journeys in the Sciences.”

[12] R. Tibshiranit, “Regression Shrinkage and Selection via the Lasso,” 1996. [Online]. Available: https://academic.oup.com/jrsssb/article/58/1/267/7027929