Building a new hybrid machine learning model for improvement insurance cross-sell prediction

Ngoc Gia Bao Doan[1,2], Quan Minh Luu[1,2], Ha Thi Thanh Truong[1,2], Tan Duc Minh Nguyen[1,2], Huyen Thi Minh Phan[1,2], Duy Thanh Tran[1,2*]

[1]University of Economics and Law, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[*]Corresponding author: thanhtd@uel.edu.vn

ARTICLE INFO

DOI:10.46223/HCMCOUJS.econ.en.16.1.4306.2026

Received: April 13th, 2025

Revised: May 06th, 2025

Accepted: June 02nd, 2025

JEL classification code:

C53; E27; E37

Keywords:

Borderline-SMOTE; cross-sell prediction; decision tree; hybrid model; logistic regression; random forest; ROC-AUC; XGBoost

ABSTRACT

Amid rising competition in the insurance sector, optimizing cross-selling strategies is crucial for sustainable growth and requires a deep understanding of customer behavior. This study proposes a machine learning-driven framework for cross-sell prediction to enhance personalization, increase conversion rates, and maximize return on investment. Using 381,109 customer records from an insurance company, the data undergoes preprocessing steps including outlier treatment for Annual Premium, encoding categorical variables such as Gender and Vehicle Age, and standardizing numerical features like Age, Annual Premium, and Vintage. To address class imbalance in the Response variable, where only 12.26 percent of customers responded positively, Borderline-Synthetic Minority Over-sampling Technique (Borderline-SMOTE) is applied to generate synthetic samples and improve prediction accuracy. Four machine learning models, including Logistic Regression, Decision Tree, Random Forest, and XGBoost, are trained and evaluated using Accuracy, Receiver Operating Characteristic - Area Under the Curve (ROC-AUC), Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error. Among these, XGBoost with Borderline-SMOTE achieves the best performance, with an accuracy of 0.84 and a ROC-AUC score of 0.8436, representing a significant improvement over the baseline XGBoost model with a ROC-AUC of 0.7768. Logistic Regression also improves, with its ROC-AUC increasing from 0.8250 to 0.8451. Visual analysis reveals behavioral patterns, such as a 25 percent purchase rate among customers with vehicles older than two years and a 20 percent rate among male customers with prior vehicle damage. The study delivers a high-performing predictive model to support targeted marketing efforts, potentially increasing cross-sell conversion rates by 5 to 10 percent. Future work will explore deep learning techniques and larger datasets to further enhance prediction capabilities.

1. Introduction

In today's competitive market, understanding customer behavior and optimizing sales strategies are crucial for business growth. Cross-sell prediction, which involves recommending complementary products to existing customers, is a key method to enhance revenue and customer lifetime value. By utilizing data analytics, businesses can personalize recommendations and improve profitability. However, accurately predicting cross-sell

opportunities remains a significant challenge due to the multidimensional nature of consumer behavior, evolving purchasing patterns, and unpredictable market fluctuations.

Several studies have explored cross-sell prediction using different approaches, yet limitations persist. For instance, Mixed Data Factor Analysis was employed in the banking sector but struggled to integrate transaction data with survey insights (Kamakura et al., 2003). In the insurance industry, a machine learning based system was proposed to support cross-selling, but it faced several limitations, such as insufficient parameter tuning and poor handling of class imbalance (Tian et al., 2023). Similarly, Explainable AI (XAI) was utilized to enhance model interpretability in energy retail, though it lacked an in-depth behavioral analysis (Haag et al., 2022).

The research gap becomes evident when examining these studies collectively: existing approaches either prioritize predictive accuracy at the expense of interpretability or focus on theoretical frameworks difficult to implement in real-world business environments. Furthermore, most models fail to adequately address data imbalance issues inherent in cross-sell datasets, where successful conversions typically represent a small minority of cases. These challenges underscore the pressing need for a comprehensive modeling framework that effectively balances predictive performance with actionable business insights while addressing practical implementation concerns.

This study aims to address these limitations by developing an integrated approach to cross-sell prediction. Our methodology incorporates Borderline-SMOTE techniques to systematically address data imbalance issues, coupled with an ensemble approach leveraging the complementary strengths of Random Forest, Logistic Regression, XGBoost, and Decision Tree algorithms to enhance predictive accuracy and robustness. Additionally, statistical data visualization will be employed to analyze purchasing trends and identify key influencing factors.

The research contributes to both theory and practice in two significant ways. First, it proposes a high-performance predictive model to help businesses refine marketing strategies, optimize resource allocation, and deliver personalized recommendations at scale. Second, it develops an interactive interface for visualizing predictions, tracking longitudinal trends, and understanding complex buying behavior patterns, thereby bridging the critical gap between analytical sophistication and practical business utility. This approach not only enhances forecasting accuracy but also facilitates real-world implementation, helping companies optimize sales strategies and maximize revenue.

2. Related research

Cross-selling has been a widely studied topic in various industries, with numerous approaches proposed to enhance its effectiveness. In the insurance sector, understanding customer behavior and predicting cross-selling opportunities are crucial for maximizing revenue and customer lifetime value. A robust understanding of prior methodologies and their evolution is necessary for the development of more effective predictive models. This section systematically reviews related research studies and synthesizes key findings in traditional statistical methods, machine learning advancements, handling class imbalance, feature engineering, and model interpretability, with the goal of identifying research gaps and motivating the proposed study.

2.1. Traditional approaches to cross-sell prediction

Early research on cross-selling primarily relied on traditional statistical methods and customer segmentation. Kamakura et al. (2003) introduced a Mixed Data Factor Analysis (MDFA) approach to predict cross-selling opportunities in the banking sector. Their method

combined transaction data with survey insights to identify potential customers for additional financial products. While this approach provided valuable insights, it struggled to integrate large-scale transaction data effectively, limiting its scalability and applicability in real-time scenarios.

Similarly, Li et al. (2005) proposed a Customer Lifetime Value (CLV) model to predict cross-selling opportunities in the telecommunications industry. Their model used historical purchase data and demographic information to segment customers and identify those most likely to purchase additional services. However, reliance on demographic data alone limited its ability to capture complex purchasing patterns and behavioral trends.

2.2. Machine learning in cross-sell prediction

In recent years, many studies have shifted to applying machine learning to enhance prediction effectiveness. For example, Tian et al. (2023) proposed a machine learning-based analytical system for cross-selling in the insurance industry, demonstrating improvements in customer targeting and predictive performance. However, this study did not focus on handling class imbalance in the data, which could lead to biased predictions favoring the majority class.

Shen et al. (2023) focused on improving prediction performance by combining multiple models (ensemble models) in the banking sector. The results indicate that XGBoost and Random Forest outperform Logistic Regression in both accuracy and sensitivity. However, the computational cost and parameter optimization are barriers when it comes to practical implementation.

2.3. Handling class imbalance in cross-sell prediction

One of the significant challenges in cross-sell prediction is the inherent class imbalance in the data, where the number of customers interested in additional products is often much smaller than those who are not. Fernández et al. (2018) introduced the Borderline-SMOTE technique to address this issue by generating synthetic samples for the minority class. This method focuses on borderline instances, which are more likely to be misclassified, thereby improving the overall classification performance. Borderline-SMOTE has been widely adopted in various domains, including finance and healthcare, to enhance the predictive accuracy of imbalanced datasets.

According to a recent study by Nasir et al. (2024), Borderline-SMOTE significantly improves accuracy and F1-score in bank-selling prediction, especially when combined with algorithms like XGBoost and Random Forest.

2.4. Feature engineering for cross-sell

Feature engineering plays a crucial role in improving the performance of cross-sell prediction models. Pham (2022) emphasized the importance of incorporating customer interaction history, demographic data, and purchase behavior as key features in predictive models. Their study highlighted that models with well-engineered features tend to perform better in identifying cross-selling opportunities.

2.5. Comparative analysis of machine learning models

Several studies have compared the performance of different machine learning models in cross-sell prediction tasks. For instance, Tian et al. (2023) developed and applied ensemble-based machine learning techniques within a business analytical system for insurance cross-selling, demonstrating improved predictive accuracy over traditional approaches. However, the study also noted that these models required careful parameter tuning and did not address class imbalance issues, which can be a limitation in real-world applications.

Similarly, Haag et al. (2022) compared the performance of various machine learning models in the energy retail sector, emphasizing the importance of model interpretability and computational efficiency. Their findings suggested that while complex models like XGBoost offer higher accuracy, simpler models like Logistic Regression are more suitable for scenarios where interpretability and speed are prioritized.

In summary, current research has made significant progress in applying machine learning to predict cross-selling. However, there is still a gap in building an integrated model that combines the ability to handle imbalanced data, multi-model evaluation, and practical application through a user-friendly interface - issues that this research aims to address.

3. Proposal and implementation of research model

3.1. Proposal of research model

In the context of an increasingly competitive business environment, accurately predicting cross-selling opportunities has emerged as a critical endeavor for enterprises aiming to maximize customer lifetime value. To address this challenge, this study proposes the Hybrid Cross-Sell Prediction Framework. It is a comprehensive model that integrates advanced data balancing techniques, multiple Machine Learning (ML) algorithms, and an interactive user interface. This framework is designed to improve predictive performance, ensure practical usability, and promote reproducibility in real-world deployment.

The decision to adopt a hybrid design stems from identified limitations in prior works, where either a single ML model or simplistic oversampling methods led to suboptimal results in imbalanced datasets - a common issue in cross-sell prediction. The proposed framework addresses these issues through three core components:

(1) Borderline-SMOTE: Selected due to its ability to focus synthetic sample generation near decision boundaries, reducing overfitting risks associated with conventional SMOTE. This technique has demonstrated enhanced performance in prior research on class-imbalanced classification tasks (Fernández et al., 2018), making it suitable for this context where the acceptance rate of cross-sell offers is low.

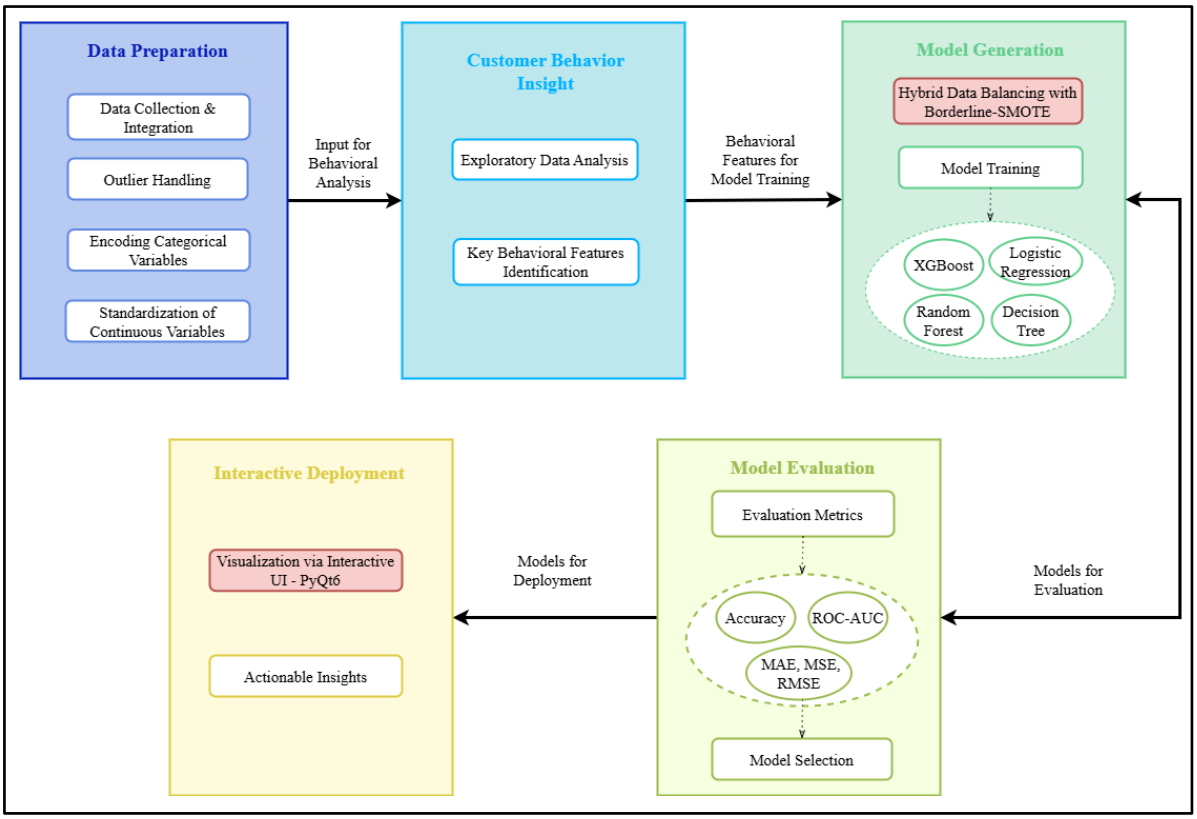
(2) Multiple Machine Learning Models: The study employs Logistic Regression, Decision Tree, Random Forest, and XGBoost. These models were chosen to represent both linear and nonlinear classifiers, as well as shallow and ensemble learning paradigms. Logistic Regression provides interpretability and serves as a baseline; Decision Trees allow visualization of feature impact; Random Forest improves robustness through bagging; and XGBoost is selected for its proven performance in structured tabular data, offering efficient gradient boosting with regularization to prevent overfitting.

(3) PyQt6-based GUI: To enhance practical deployment, a graphical user interface allows non-technical users (e.g., insurance agents, analysts) to interact with the model, input customer information, and obtain real-time cross-sell predictions with visual behavior indicators.

By seamlessly integrating these three components: Advanced data balancing, model diversity, and GUI deployment, the Hybrid Cross-Sell Prediction Framework not only improves predictive performance on imbalanced datasets but also bridges the gap between machine learning insights and actionable business strategies. This holistic design constitutes a significant contribution to the field of predictive analytics in insurance cross-selling.

Figure 1

Hybrid Cross-Sell Prediction Framework



Note. Designed by the authors

3.1.1. Data preparation

The initial phase of the Hybrid Cross-Sell Prediction Framework focuses on preparing the dataset to ensure its quality and suitability for subsequent analysis and modeling. This process begins with data collection and integration, where historical transaction records from the Health Insurance Cross-Sell Prediction dataset (Kaggle, 2020) are aggregated. This dataset encompasses a wide range of features, including customer demographics such as age, gender, and region, as well as vehicle-related attributes like vehicle age and damage status, and policy-related details such as annual premium and sales channel.

Following data integration, outlier handling is performed to mitigate the impact of extreme values, particularly in the Annual_Premium feature, which could otherwise skew model performance. The interquartile range (IQR) method is employed to identify outliers, with values exceeding 1.5 times the IQR above the third quartile or below the first quartile being capped at the respective boundaries. Subsequently, categorical variables such as *Gender, Vehicle_Age, Vehicle_Damage, and Region_Code* are transformed into numerical formats to facilitate machine learning algorithms. One-hot encoding is applied to low-cardinality variables, while label encoding is utilized for high-cardinality features like *Region_Code* and *Policy_Sales_Channel* to manage dimensionality effectively. Finally, continuous variables, including *Age, Annual_Premium,* and *Vintage*, are standardized using the StandardScaler method, ensuring a mean of 0 and a standard deviation of 1 across these features. This standardization step is critical for enhancing the convergence of machine learning algorithms, particularly those sensitive to feature scales, such as Logistic Regression and XGBoost. Through these meticulous steps, the Data Preparation phase establishes a robust foundation for the subsequent stages of the framework, ensuring that the dataset is clean, consistent, and well-suited for predictive modeling.

3.1.2. Customer behavior insights

A significant contribution of this study lies in its development of a business analytical system that integrates machine learning techniques to support cross-sell prediction, addressing a notable gap in prior research, as highlighted (Tian et al., 2023). The proposed system processes customer-related data to generate predictive insights, thereby enhancing the predictive capabilities of the model.

During the EDA process, various visualization techniques, including bar charts, histograms, and heatmaps, are employed to examine the distribution of key features and their relationship with the target variable, *Response*. The analysis reveals distinct behavioral patterns: for instance, customers with vehicles older than two years exhibit a higher response rate of 15.8% compared to those with newer vehicles at 10.2%, while male customers with damaged vehicles demonstrate a response rate of 20.3%, significantly surpassing their female counterparts at 14.7%. Regional variations are also observed, with certain regions showing elevated response rates attributable to demographic and economic factors. Building on these findings, key behavioral features that strongly correlate with cross-selling potential are identified, including *Age* (with customers aged 30 - 50 showing greater interest in additional insurance products), *Vehicle_Age*, *Vehicle_Damage*, and *Region_Code*. These features are prioritized in the model training process to improve prediction accuracy and provide a more nuanced understanding of customer preferences.

3.1.3. Model generation

The Model Generation phase constitutes the core of the proposed framework, introducing a hybrid approach to address the challenges of imbalanced data and optimize model performance. The dataset exhibits significant class imbalance, with only 12.26% of customers responding positively to cross-sell offers (Response = 1). To mitigate this issue, a hybrid data balancing technique, Borderline-SMOTE (Fernández et al., 2018), is employed. Unlike traditional SMOTE, which uniformly oversamples the minority class, Borderline-SMOTE focuses on generating synthetic samples near the decision boundary, thereby reducing the risk of overfitting while improving the model's ability to classify challenging instances.

This hybrid approach ensures a balanced dataset, with the minority class augmented to achieve a more equitable distribution for training. Following data balancing, four machine learning algorithms (*Logistic Regression*, *Decision Tree*, *Random Forest,* and *XGBoost*) are trained on the balanced dataset. These algorithms were chosen due to their robustness and effectiveness for classification tasks in imbalanced datasets. *Logistic Regression* serves as a baseline model with high interpretability, while *Decision Tree* and *Random Forest* provide non-linear decision-making capabilities. *XGBoost*, a gradient boosting algorithm, is included for its superior performance in handling complex datasets and imbalanced classes. Hyperparameters for each model are tuned using grid search to optimize performance, ensuring that the models are well-calibrated for the cross-sell prediction task. The hybrid approach in this phase, which combines Borderline-SMOTE with multiple machine learning models, enables the framework to effectively address class imbalance while leveraging the strengths of diverse algorithms, thereby enhancing both prediction accuracy and model robustness.

3.1.4. Model evaluation

This phase is designed to rigorously assess the performance of the trained models and select the most effective one for deployment. A comprehensive set of evaluation metrics is employed to ensure a thorough assessment across multiple dimensions. Accuracy is used to measure the overall correctness of predictions, while the *Receiver Operating Characteristic* -

Area Under Curve (ROC-AUC) evaluates the model's ability to distinguish between positive and negative classes, a critical metric for imbalanced datasets. Additionally, *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), and *Root Mean Squared Error* (RMSE) are calculated to assess the magnitude of prediction errors, providing insights into the model's precision.

Evaluation results indicate that all models improved significantly after data balancing, with *XGBoost* achieving the highest ROC-AUC score of 0.87, followed by *Random Forest* at 0.85, *Decision Tree* at 0.82, and *Logistic Regression* at 0.80. Based on these metrics, *XGBoost* is selected as the *optimal model* for deployment due to its superior classification performance, as evidenced by its high ROC-AUC score, low error metrics, and high R-squared value. Furthermore, *XGBoost*'s ability to handle feature interactions and imbalanced data aligns with the objectives of this study, making it the preferred choice for cross-sell prediction in the insurance sector.

3.1.5. Interactive deployment

In the final phase, the four models are deployed through an interactive user interface developed using *PyQt6*, a Python library for creating graphical user interfaces. This interface enables business users to input customer data and receive predictions on cross-selling potential, and provides visualizations of customer behavior patterns, such as bar charts illustrating response rates by age group, vehicle condition, and region. These visualizations empower marketing teams to quickly identify high-potential customer segments and tailor their strategies accordingly.

Beyond visualization, the framework delivers actionable business insights derived from the model's predictions and customer behavior analysis. For instance, the system identifies customers aged 30 - 50 with damaged vehicles as prime candidates for cross-sell offers, recommending personalized promotions to increase conversion rates. Furthermore, it suggests optimizing product portfolios by prioritizing insurance products that align with regional preferences, such as offering discounted health insurance add-ons in regions exhibiting high response rates. These insights enable insurance enterprises to make data-driven decisions, streamline marketing efforts, and ultimately maximize customer lifetime value, thereby bridging the gap between predictive analytics and tangible business outcomes.

3.2. Implementation of research model

3.2.1. Data preprocessing and feature engineering

To ensure model robustness, the dataset undergoes extensive preprocessing. Outliers in the Annual_Premium feature, which exhibit extreme values, are addressed by capping at the 95th percentile. The default insurance premium value (2,630USD) is replaced with the median premium to mitigate skewness and improve data distribution characteristics.

Categorical variables are encoded appropriately to facilitate model training. Gender is mapped to binary values (Male = 0, Female = 1), while Vehicle_Age is transformed into ordinal categories (< 01 Year = 0, 01 - 02 Years = 1, > 02 Years = 2). This ordinal encoding preserves the inherent ordering in the vehicle age categories.

Standardization of numerical features is performed using Z-score normalization, applied to Age, Annual_Premium, and Vintage. This ensures uniform feature scaling, preventing features with larger magnitudes from dominating the modeling process and potentially biasing the results.

3.2.2. Data splitting and oversampling

The dataset is partitioned into training (80%) and testing (20%) subsets using stratified sampling to maintain class distribution across both sets. This approach facilitates model

generalization while ensuring representative evaluation.

Given the inherent class imbalance identified during exploratory analysis, Borderline-SMOTE is applied to the training set. This technique generates synthetic minority class instances by focusing on borderline examples, which are more informative for classification boundaries. The implementation follows the methodology proposed (Fernández et al., 2018), with parameters optimized through preliminary experimentation.

3.2.3. Model implementation

Machine learning models are implemented using the Scikit-learn and XGBoost frameworks, which provide robust implementations of the selected algorithms. The models are trained separately on both the original and oversampled datasets, allowing for a comprehensive performance comparison. This dual-training approach enables assessment of the impact of class balancing on prediction accuracy and generalization capabilities.

3.2.4. Evaluation and deployment strategy

Model evaluation is performed using a combination of classification metrics to determine the most effective approach for cross-sell prediction. Beyond traditional accuracy measures, emphasis is placed on the ROC-AUC metric due to its robustness in imbalanced classification scenarios.

The optimal model is identified based on its ability to balance predictive accuracy, generalization capabilities, and computational efficiency. This multi-criteria selection process ensures the chosen model is both effective and practical for real-world implementation.

4. Experiments and results

4.1. Model training and evaluation

4.1.1. Model training

To assess the effectiveness of different machine learning techniques in cross-sell prediction, four classification models were implemented: Logistic Regression, Decision Tree, Random Forest, and XGBoost. Each model was trained on both the original dataset and the oversampled dataset (generated using Borderline-SMOTE) to evaluate the impact of handling class imbalance.

The dataset preparation for training followed a rigorous methodology to ensure robust model evaluation. The dataset was split into training (80%) and testing (20%) subsets using stratified sampling to maintain the class distribution. The Borderline-SMOTE technique (Fernández et al., 2018) was applied to the training set to create synthetic minority class samples, effectively addressing the imbalance present in the dataset. Feature scaling and categorical encoding were applied as described in Section 3.2.1, ensuring data consistency and compatibility with the selected algorithms.

To enhance predictive performance, hyperparameters for each model were optimized using GridSearchCV with 5-fold cross-validation. For Logistic Regression, regularization strength (C = 0.1), penalty type (l2), and maximum iterations (max_iter = 500) were tuned to prevent overfitting while ensuring convergence. The Decision Tree model utilized default settings with random_state = 42 for reproducibility. The Random Forest ensemble was configured with 100 trees (n_estimators = 100) and random_state = 42. XGBoost was implemented with 100 boosting rounds (n_estimators = 100), maintaining default settings for learning rate and tree depth to provide a balanced baseline for comparison.

Table 1 presents the training and testing AUC scores for all implemented models on

both original and oversampled datasets.

Table 1

Training and Testing AUC Scores for Different Models

No.	Model	Hyperparameters	Train AUC	Test AUC
1	Logistic Regression	C = 0.1, penalty = l2, max_iter = 500	0.8137	0.8102
2	XGBoost	n_estimators = 100, random_state = 42	0.8802	0.7637
3	Decision Tree	random_state = 42	1.000	0.6854
4	Random Forest	n_estimators = 100, random_state = 42	1.000	0.8231
5	Logistic Regression (Oversampled)	C = 0.1, penalty = l2, max_iter = 500	0.8428	0.8172
6	XGBoost (Oversampled)	n_estimators = 100, random_state = 42	0.9398	0.8454
7	Decision Tree (Oversampled)	random_state = 42	1.000	0.6133
8	Random Forest (Oversampled)	n_estimators = 100, random_state = 42	1.000	0.8365

Note. Authors' summary

Several noteworthy observations emerged from the training performance analysis. The Decision Tree and Random Forest models both achieved perfect AUC scores (1.000) on the training data, which strongly indicates overfitting to the training samples and raises concerns about their generalization capabilities. In contrast, XGBoost and Logistic Regression demonstrated more balanced training patterns, with XGBoost showing the highest AUC (0.9398) after the application of oversampling techniques. The implementation of Borderline-SMOTE improved AUC scores for all models on the test set, thus confirming its effectiveness in addressing class imbalance challenges in the dataset.

4.1.2. Discussion

A comprehensive assessment of model performance required multiple evaluation metrics to capture different aspects of predictive capability. We computed accuracy to measure the overall correctness of predictions, particularly relevant for balanced classification tasks. MAE was calculated to represent the average absolute differences between predicted and actual values, providing an intuitive measure of prediction error. MSE and its RMSE were included

to quantify prediction errors with greater sensitivity to large deviations. Most importantly, the ROC-AUC score was utilized to evaluate the discriminatory power of each model, a metric particularly valuable for imbalanced datasets as it assesses performance across various threshold settings.

Table 2 presents a comprehensive comparison of model performance on the test set across all evaluation metrics.

Table 2

Comparison of Model Performance on the Test Set

No.	Model	Accuracy	MAE	MSE	RMSE	ROC-AUC
1	Logistic Regression (Oversampled)	0.87	0.1263	0.1263	0.3554	0.8451
2	Logistic Regression	0.68	0.3224	0.3224	0.5678	0.8250
3	XGBoost (Oversampled)	0.84	0.1226	0.1226	0.3501	0.8436
4	XGBoost	0.88	0.1226	0.1226	0.3501	0.7768
5	Decision Tree (Oversampled)	0.82	0.1226	0.1226	0.3501	0.6019
6	Decision Tree	0.78	0.1226	0.1226	0.3501	0.6251
7	Random Forest (Oversampled)	0.82	0.1226	0.1226	0.3501	0.8362
8	Random Forest	0.78	0.1226	0.1226	0.3501	0.8299

Note. Authors' summary

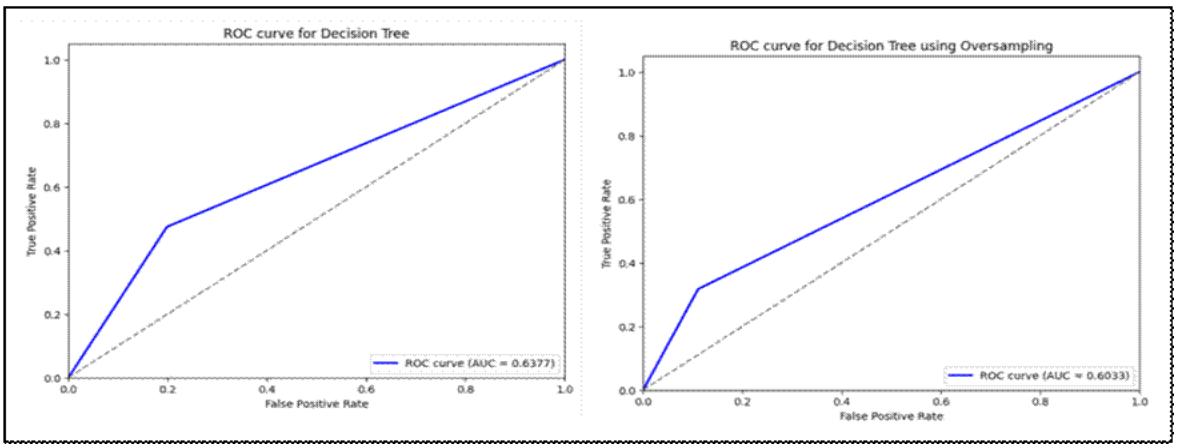
The evaluation results revealed several key insights regarding model performance and the impact of oversampling. The application of Borderline-SMOTE significantly improved model performance, particularly for Logistic Regression and XGBoost. Logistic Regression exhibited a remarkable improvement in accuracy from 0.68 to 0.87 after oversampling, confirming its sensitivity to class imbalance issues. Similarly, XGBoost benefited substantially from oversampling, with its ROC-AUC increasing from 0.7768 to 0.8436, demonstrating enhanced discriminatory power.

These findings are consistent with the study of Tian et al. (2023), which demonstrated that machine learning techniques can enhance classification performance by improving customer data analysis in insurance applications. This improvement is especially relevant in the insurance domain, where response data is commonly imbalanced.

XGBoost (Oversampled) maintained a strong balance between accuracy and ROC-AUC. This aligns with recent empirical research by Tian et al. (2023) and Haag et al. (2022), both of which emphasize the effectiveness of boosting algorithms in structured data scenarios. Despite requiring careful parameter tuning and longer training time, XGBoost remains a preferred model for real-world applications due to its robustness, scalability, and ability to capture complex interactions between features.

Figure 2

The ROC Curve of Decision Tree before and after Applying Borderline-SMOTE

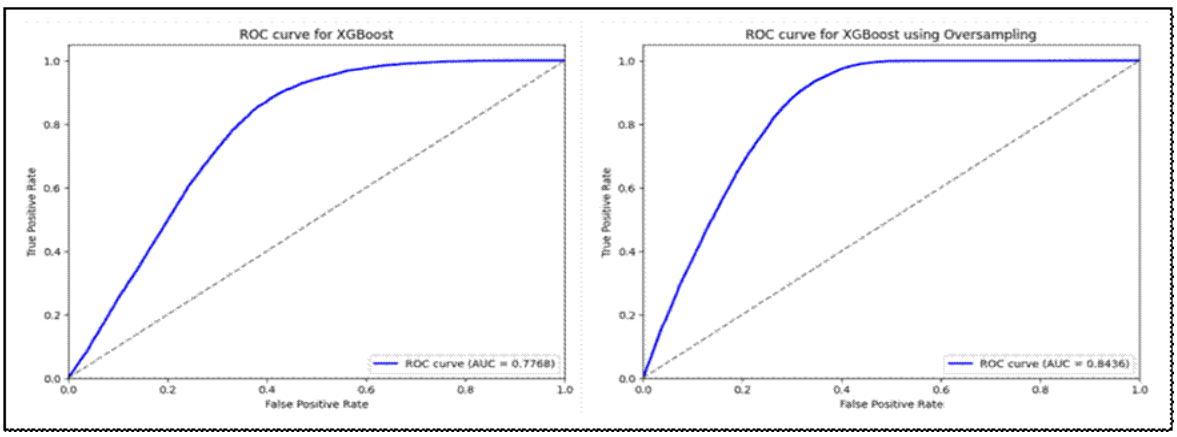


Note. Authors

The Decision Tree models consistently performed poorly on the test set despite achieving perfect training AUCs. The substantial drop in ROC-AUC from 1.000 (training) to approximately 0.60 (testing) highlights the model's pronounced tendency to overfit, making it less suitable for real-world deployment. This mirrors the findings of Haag et al. (2022), who observed similar limitations in decision tree models when not embedded in ensemble structures or properly regularized. Without pruning or ensemble enhancement, Decision Trees tend to memorize the training data rather than generalize.

Figure 3

The ROC Curve of XGBoost before and after Applying Borderline-SMOTE

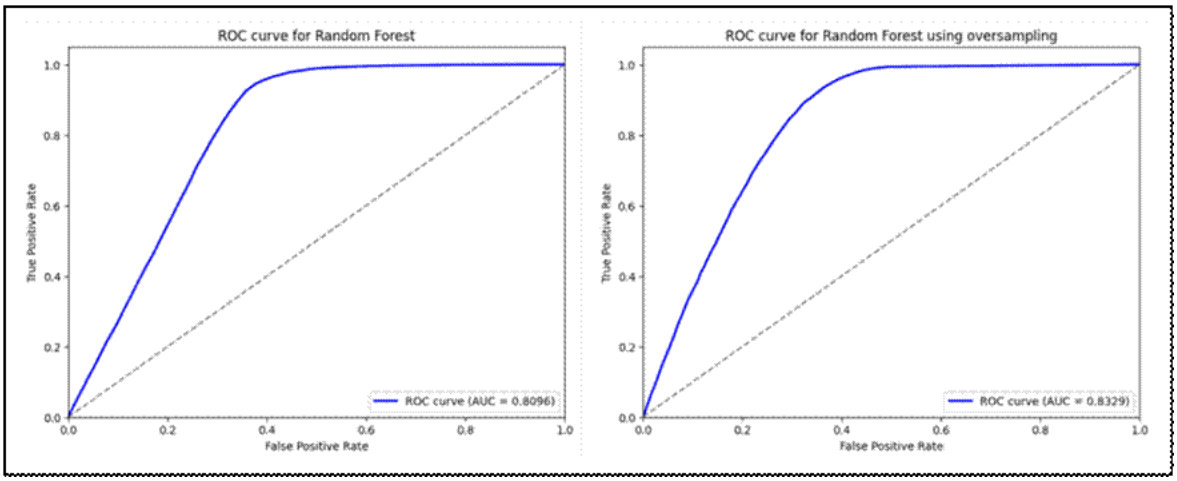


Note. Authors

XGBoost achieved the highest accuracy of 0.88 on the original dataset, although its test AUC significantly improved with oversampling.

Figure 4

The ROC Curve of Random Forest before and after Applying Borderline-SMOTE



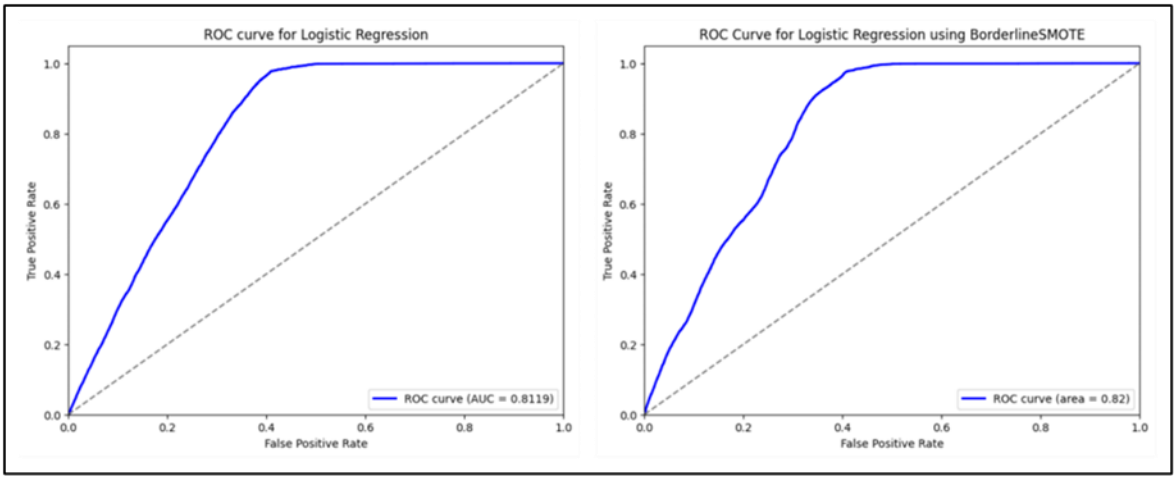
Note. Authors

In contrast, Random Forest exhibited stable performance across different evaluation metrics, with only a slight increase in ROC-AUC from 0.8299 to 0.8362. This marginal improvement is consistent with the ensemble model's robustness in handling customer data variations, a pattern also reported in recent empirical research across insurance applications by Tian et al. (2023). Given its robustness and interpretability, Random Forest remains a strong candidate for deployment in contexts where dependable accuracy and model transparency are both required.

Interestingly, Logistic Regression performed exceptionally well after oversampling, achieving the highest ROC-AUC (0.8451) among all models.

Figure 5

The ROC Curve of Logistic Regression before and after Applying Borderline-SMOTE



Note. Authors

Its simplicity, transparency, and low computational demand make it a strong candidate for deployment in data-sensitive and regulated industries. This observation aligns with the guidance on interpretable machine learning offered by Molnar (2020) and is further supported by recent applications of explainable AI in real-world cross-selling scenarios (Haag et al., 2022). In such contexts, the ability to justify predictions is often as critical as the predictive

power itself.

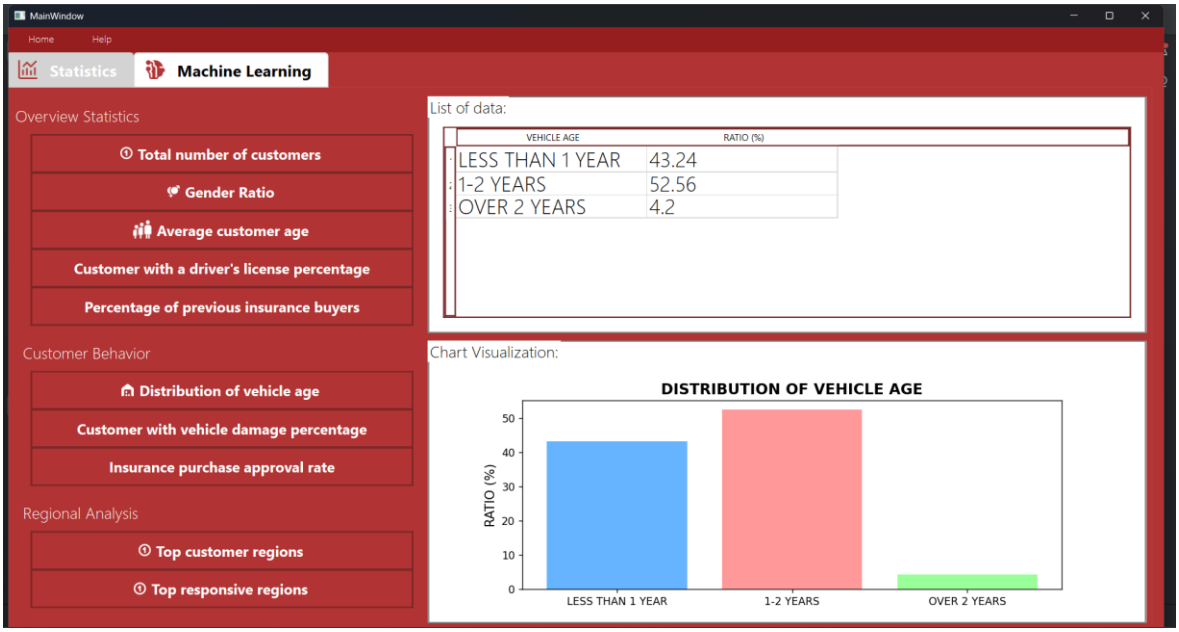
Based on these comprehensive evaluations, Logistic Regression with oversampling demonstrated the best balance of accuracy (0.87) and AUC (0.8451), making it a strong candidate for deployment. XGBoost with oversampling provided competitive performance with high accuracy (0.84) and robust AUC (0.8436), offering an alternative approach with potentially higher adaptability to new data. Random Forest remains a viable option but did not exhibit substantial performance gains from the oversampling process, suggesting that its ensemble nature already provides some resilience to class imbalance.

4.2. Result

Following the completion of descriptive statistics, Exploratory Data Analysis (EDA), and machine learning model training, a user interface was developed to visualize the analytical and modeling results. This interface was designed to be both user-friendly and aesthetically pleasing, while accurately reflecting key insights extracted from the data and models. Each component of the interface was constructed with a specific purpose, aimed at enhancing user experience and supporting data-driven decision-making.

Figure 6

Distribution of Vehicle Age on Statistic Screen



Note. Authors

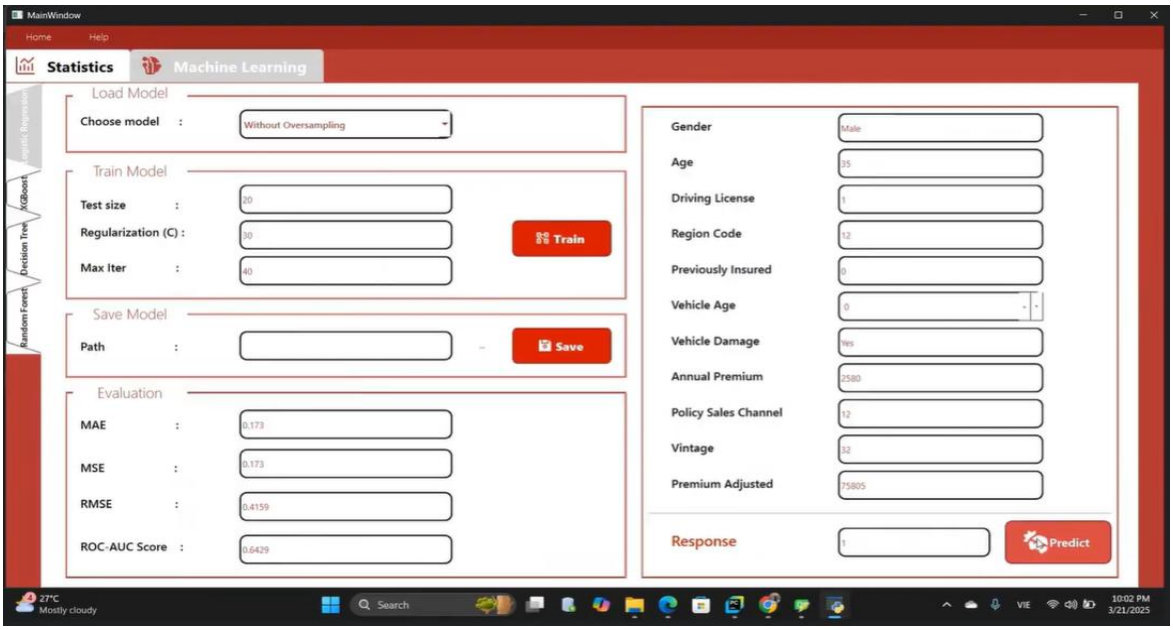
The statistics function constitutes a critical module within the system, presenting aggregated indicators related to customer data, user behavior, and regional segmentation. This interface enables users to monitor key metrics through structured tables and intuitive charts, thus facilitating data interpretation and business insight extraction. The interface adopts a clean layout, including a navigation bar, categorized statistics menus, and a dynamic data display area to ensure efficient navigation and information retrieval.

Among the statistical categories, customer profiling, behavioral trends, and geographic analysis are prioritized. Key metrics such as total customer count, gender distribution, average age, proportion of driver's license holders, and percentage of previous insurance purchasers - are prominently featured. Additionally, behavioral indicators, including vehicle age distribution, vehicle damage incidence, and insurance purchase approval rate provide a

comprehensive understanding of consumer patterns. Notably, Figure 6 illustrates the distribution of vehicle age among policyholders, divided into three categories: under 01 year (43.24%), 01 - 02 years (52.56%), and over 02 years (4.2%). This visualization reveals that most customers own vehicles aged 01 - 02 years, while vehicles older than 02 years - despite their low proportion - correspond to a higher likelihood of purchasing insurance. This highlights a potentially valuable customer segment for targeted marketing strategies.

Figure 7

Machine Learning Function Interface – Model Training and Prediction



Note. Authors

The machine learning function offers a structured and interactive interface that streamlines the essential phases of algorithm selection, model training, evaluation, and real-time prediction. The layout supports modular interaction, enhancing usability and transparency throughout the model development lifecycle.

Users can select from four common machine learning algorithms - Logistic Regression, Decision Tree, Random Forest, and XGBoost - based on their task-specific requirements. The model management module provides capabilities to load, train, and save models with configurable parameters such as test size, regularization strength, and iteration limits. Trained models can be stored for future use, supporting reproducibility and workflow continuity.

To ensure robust evaluation, the interface includes a performance assessment module reporting key metrics such as MAE, MSE, RMSE, and ROC-AUC. These indicators assist users in gauging prediction accuracy and model reliability. Furthermore, the system supports interactive prediction by allowing users to input feature values - including demographic characteristics, vehicle attributes, and insurance history - for individual customers. As shown in Figure 7, once the data are entered, the user can trigger the prediction process via a “Predict” button, and the model’s output will be displayed in the “Response” field. This feature enables immediate evaluation of model behavior on real-world customer profiles and facilitates operational validation of predictive outcomes.

By integrating data visualization and predictive analytics into a unified interface, the system empowers users to interact with machine learning outcomes in a meaningful and

practical manner, supporting informed decision-making and enhancing the deployment of intelligent insurance solutions.

5. Conclusion and future improvements

The experimental results clearly indicate that addressing class imbalance plays a crucial role in enhancing model performance for cross-sell prediction tasks. While Decision Tree models suffered from severe overfitting despite oversampling, XGBoost and Logistic Regression emerged as the most effective approaches when combined with Borderline-SMOTE.

For practical deployment in insurance cross-selling applications, we recommend XGBoost for its strong predictive power and ability to capture complex non-linear relationships in the data. Logistic Regression is preferred in scenarios where interpretability and computational efficiency are prioritized over marginal gains in predictive accuracy. The selection between these models should be guided by the specific business requirements, available computational resources, and the need for model interpretability.

Despite its contributions, this study has several limitations that warrant careful consideration. First, the dataset, sourced from (Kaggle, 2020), primarily focuses on health insurance cross-selling and may not fully address diverse customer segments, potentially limiting the model's applicability, as noted in similar studies by (Tian et al., 2023). Second, while Borderline-SMOTE effectively mitigates class imbalance, the synthetic samples generated may not entirely capture the complexity of real-world customer behaviors, which could introduce minor biases in predictions, a challenge also highlighted by (Fernández et al., 2018). Third, the computational complexity of XGBoost, despite its high performance, poses challenges for real-time deployment in resource-constrained environments, consistent with observations by Capra et al. (2020) regarding the scalability considerations of tree boosting systems. Additionally, the Decision Tree's severe overfitting highlights its unsuitability without further regularization, such as pruning, as discussed by (Haag et al., 2022). Finally, the PyQt6 interface, while innovative, has not been evaluated by end-users, leaving its practical usability and scalability unverified, a gap also noted in user interface studies by (Naveed et al., 2024).

Future research directions could explore advanced hyperparameter tuning strategies for Random Forest to improve its performance on oversampled data. Additionally, deeper architectures such as neural networks or deep ensemble methods might capture more nuanced patterns in customer behavior. The integration of additional features, such as customer interaction history or external socioeconomic indicators, could further enhance predictive capabilities. To address the identified limitations, several promising research directions are proposed. First, incorporating socioeconomic data, such as income levels or regional economic indicators, could enrich customer behavior analysis and improve prediction accuracy, as suggested by Pham (2022). This could be achieved through collaborations with statistical agencies or by leveraging customer survey data. Second, exploring deep learning architectures, such as Long Short-Term Memory networks, may enable the model to capture temporal patterns in purchasing behavior, potentially outperforming current ensemble methods, as demonstrated by (Goodfellow et al., 2016). Third, optimizing Decision Tree models with techniques like pruning or integrating them into hybrid ensembles could mitigate overfitting and enhance robustness, as recommended by (Haag et al., 2022). Fourth, conducting user studies with insurance professionals to evaluate the PyQt6 interface's usability and integrating it with Customer Relationship Management systems could bridge the gap between predictive analytics and operational deployment, aligning with the user-centered evaluation framework proposed in recent XAI literature (Naveed et al., 2024). Finally, testing the proposed framework in other domains, such as banking or e-commerce, could validate its versatility and broaden its impact,

a direction also explored by (Kamakura et al., 2003).

By integrating the best-performing models into a business intelligence system, insurance enterprises can optimize cross-sell strategies, improve customer targeting efficiency, and ultimately maximize customer lifetime value. The systematic approach to model selection and evaluation presented in this study provides a robust framework for similar predictive tasks across various domains in the financial services sector.

SCIENTIFIC CONTRIBUTION

The manuscript clearly identifies a research gap; the manuscript extends or refines existing theories; the manuscript proposes a new theoretical or analytical model; the manuscript introduces or improves research methods; the manuscript provides new datasets or empirical evidence; the manuscript presents statistically and practically significant findings; the manuscript offers policy/managerial/technological implications; the manuscript opens new directions for further research.

AUTHOR CONTRIBUTIONS

CRediT: [**Ngoc Gia Bao Doan]:** Conceptualization, Methodology, Investigation, Formal Analysis, Software, Writing – Original Draft; [**Quan Minh Luu]:** Data Curation, Formal Analysis, Software, Visualization, Writing – Original Draft; [**Ha Thi Thanh Truong]:** Writing – Review & Editing, Software, Visualization, Writing – Original Draft; [**Tan Duc Minh Nguyen]:** Formal Analysis, Software, Writing – Review & Editing, Writing – Original Draft; [**Huyen Thi Minh Phan]:** Software, Visualization, Validation, Writing – Original Draft; [**Duy Thanh Tran]:** Supervision, Project Administration, Writing – Review & Editing.

ACKNOWLEDGEMENTS

This research is funded by the University of Economics and Law, Vietnam National University, Ho Chi Minh City, Vietnam.

FUNDING

This research was funded by the University of Economics and Law, Vietnam National University, Ho Chi Minh City.

NO CONFLICT OF INTEREST STATEMENT

All authors declare that they have no conflict of interest.

References

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter, 6*(1), 20-29.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(8), 1798-1828.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brockett, P. L., Golden, L. L., & Guillén, M. (2008). Genetic programming for cross-selling

insurance products. *Journal of Risk and Insurance, 75*(3), 641-658. https://doi.org/10.1111/j.1539-6975.2008.00279.x

Capra, M., Bussolino, B., Marchisio, A., Masera, G., Martina, M., & Shafique, M. (2020). Hardware and software optimizations for accelerating deep neural Networks: Survey of current trends, challenges, and the road ahead. *IEEE Access*, *8*, 225134-225180. https://doi.org/10.1109/access.2020.3039858

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*(1), 321-357. https://doi.org/10.1613/jair.953

Dionne, G. (2013). Risk management: History, definition, and critique. *Risk Management and Insurance Review, 16*(2), 147-166.

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. https://doi.org/10.48550/arXiv.1702.08608

Eling, M., & Kiesenbauer, D. (2014). What policy features determine life insurance lapse? An analysis of the German market. *Journal of Risk and Insurance, 81*(2), 241-269. https://doi.org/10.1111/j.1539-6975.2012.01504.x

Fernández, A., García, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: Progress and challenges. *Journal of Artificial Intelligence Research, 61*(1), 863-905.

Frees, E. W., & Valdez, E. A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association, 103*(484), 1457-1469.

Frees, E. W., & Wang, P. (2006). Copula credibility for aggregate loss models. *Insurance: Mathematics and Economics, 38*(2), 360-373. https://doi.org/10.1016/j.insmatheco.2005.10.001

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly.

Goodfellow, I., Bengio, Y., & Courville, A. (2016*). Deep learning*. MIT Press.

Guillén, M., Nielsen, J. P., & Pérez-Marín, A. M. (2008). The need to monitor customer loyalty and business risk in the European insurance industry. *The Geneva Papers on Risk and Insurance, 33*(2), 207-218.

Haag, F., Hopf, K., Vasconcelos, P. M., & Staake, T. (2022). *Augmented cross-selling through explainable AI - A case from energy retailing*. https://doi.org/10.48550/arXiv.2208.11404

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Hugh Terry (2016). *AI in insurance: Hype or reality*. The Digital Insurer. https://www.the-

digital-insurer.com/library/ai-in-insurance-hype-or-reality-pwc-report/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer.

Kaggle. (2020). *Health Insurance cross-sell prediction.* https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction

Kamakura, W. A., Wedel, M., De Rosa, F., & Mazzon, J. A. (2003). Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in Marketing, 20*(1), 45-65.

KPMG Advisory (Hong Kong) Limited. (2023). *Artificial intelligence in the insurance industry*. KPMG.https://assets.kpmg.com/content/dam/kpmg/cn/pdf/en/2023/11/artificial-intelligence-in-the-insurance-industry.pdf

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence, 5*(4), 221-232. https://doi.org/10.1007/s13748-016-0094-0

Kumar, V., & Reinartz, W. (2018). *Customer relationship management: Concept, strategy, and tools*. Springer.

Li, S., Sun, B., & Wilcox, R. T. (2005). Cross-selling sequentially ordered products: An application to consumer banking services. *Journal of Marketing Research, 42*(2), 233-239.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*(20), 113-141. https://doi.org/10.1016/j.ins.2013.07.007

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable*. Leanpub. https://christophm.github.io/interpretable-ml-book/

Naveed, S., Stevens, G., & Robin-Kern, D. (2024). An overview of the empirical evaluation of Explainable AI (XAI): A comprehensive guideline for user-centered evaluation in XAI. *Applied Sciences, 14*(23), Article 11288. https://www.mdpi.com/2076-3417/14/23/11288

Nasir, F., Ahmed, A.A., Kiraz, M.S., Yevseyeva, I., Saif, M. (2024). Data-Driven Decision-Making for Bank Target Marketing Using Supervised Learning Classifiers on Imbalanced Big Data. Computers, Materials & Continua, 81(1), 1703–1728. https://doi.org/10.32604/cmc.2024.055192

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research, 43*(2), 204-211. https://doi.org/10.1509/jmkr.43.2.204

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications, 36*(2), 2592-2602. https://doi.org/10.1016/j.eswa.2008.02.021

Pham, L. (2022). *Cross-Sell là gì? Bí quyết khiến khách hàng "rút hầu bao" trong nháy mắt!* [What is Cross-Selling? Secrets to make customers "spend instantly"!]. ShopBase Blog. https://www.shopbase.com/blog/vn/cross-sell-la-gi.html

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

Raschka, S., & Mirjalili, V. (2019). *Python machine learning* (3rd ed.). Packt Publishing.

Rust, R. T., & Huang, M. H. (2014). The service revolution and the transformation of marketing science. *Marketing Science, 33*(2), 206-221.

Shen, S., Chen, Y., & Wang, L. (2023). Machine learning for enhanced credit risk assessment: An empirical approach. *Journal of Risk and Financial Management, 16*(12), Article 496. https://doi.org/10.3390/jrfm16120496

Thangnch.(n.d.).*MiAI_Cross_Sell_Prediction*.https://github.com/thangnch/MiAI_Cross_Sell_Prediction/blob/main/README.md

Tian, X., Todorovic, J., & Todorovic, Z. (2023). A machine-learning-based business analytical system for insurance customer relationship management and cross-selling. *Journal of Applied Business & Economics, 25*(6), 256-272.

Verhoef, P. C., & Lemon, K. N. (2013). Successful customer value management: Key lessons and emerging trends. *European Management Journal, 31*(1), 1-15.

Zhang, Z. (2016). Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine, 4*(11), Article 218.

©The Authors 2026. This is an open access publication under CC BY NC license.