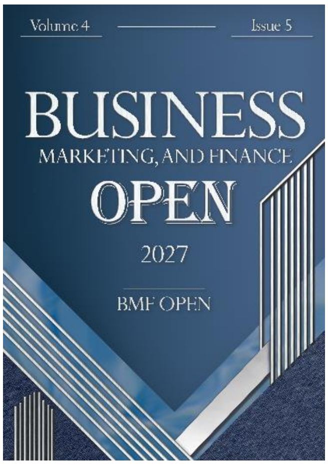
Factors Affecting Policyholders' Risk Aversion in the Adoption of Additional Endowment Life Insurance Using a Machine-Learning Algorithm

Behzad Bahmanyar [1], Hamid Khajeh Mahmoud Abadi [2,*], Ali Reza Rayati Shavazi [3] and Gholam Reza Askarzadeh [4]

[1] Department of Finance, Ya.C., Islamic Azad University, Yazd, Iran;
[2] Department of Finance, Ya.C., Islamic Azad University, Yazd, Iran;
[3] Department of Finance, Ya.C., Islamic Azad University, Yazd, Iran;
[4] Department of Finance, Ya.C., Islamic Azad University, Yazd, Iran;

* Correspondence: Ha.kha@iau.ac.ir



Citation: Bahmanyar, B., Khajeh Mahmoud Abadi, H., Rayati Shavazi, A. R., & Askarzadeh, G. R. (2027). Factors Affecting Policyholders' Risk Aversion in the Adoption of Additional Endowment Life Insurance Using a Machine-Learning Algorithm. *Business, Marketing, and Finance Open*, 4(5), 1-21.

Received: 07 April 2026
Revised: 13 July 2026
Accepted: 20 July 2026
Initial Publication: 26 July 2026
Final Publication: 01 September 2027

Abstract: This study employed machine-learning analysis on a large-scale real-world dataset comprising first and second endowment life insurance policies issued by Sarmad Insurance Company, which is recognized as a leading provider of second life insurance policies in Iran, as well as data from Fanavaran Company, an insurance technology platform operating within the insurance industry. The study population consisted of 95,962 individuals who collectively purchased their first life insurance policies through a well-known insurance company in Iran, with the policyholder and the insured being the same person. Following customer relationship management activities and follow-up contacts, 18,604 of these individuals purchased a second life insurance policy. The data were collected between 2021 and 2025 and, after data cleaning and normalization, were used as inputs for the predictive model. To predict which customers were inclined to purchase additional insurance coverage, a predictive model was developed using one of the most effective machine-learning techniques, namely Extreme Gradient Boosting (XGBoost). The predictive accuracy of the selected model was subsequently evaluated using k-fold cross-validation. Given that previous studies have consistently confirmed the risk-averse tendencies of individuals who purchase life insurance, those who purchased additional life insurance under highly unstable economic conditions—despite having the option to allocate their funds to alternative assets—may be regarded as a particularly salient group of risk-averse individuals. The model incorporated 13 demographic, behavioral, and socioeconomic factors as independent variables, while the purchase of additional endowment life insurance was considered the dependent variable. Consequently, the distinctive characteristics of customers predicted to purchase a second endowment life insurance policy were identified. The findings demonstrated significant associations between the independent variables and the probability of purchasing additional insurance. Notably, 6 of the 13 independent variables exerted a substantial influence on the purchase of additional insurance. These variables included age, gender, city and province of residence, marital status, possession of other insurance policies, and possession of automobile insurance, including third-party liability and comprehensive motor insurance. By contrast, variables such as occupation, income, and educational attainment were identified as having limited predictive influence because of the high degree of similarity among the data obtained from employees of the large organizations comprising a substantial proportion of the insurance company's customer base.

Keywords: machine-learning algorithm, endowment life insurance, additional endowment life insurance, determinants of risk aversion

1. Introduction

Risk constitutes the central organizing concept of insurance because insurance products are designed to transfer, pool, price, and manage uncertain future losses. At the individual level, the decision to purchase insurance reflects an assessment of the probability and consequences of adverse events, the affordability of premiums, confidence in the insurer, and the degree to which the individual prefers certainty over exposure to financial uncertainty. At the organizational level, insurers must simultaneously manage underwriting risk, investment risk, operational risk, liquidity risk, technological risk, and reputational risk while maintaining profitability and solvency. Consequently, risk management is not merely a technical function within insurance companies; it is a strategic capability that influences organizational performance, customer confidence, product development, and long-term competitiveness. Previous research has shown that systematic risk-management practices can improve the performance of insurance companies by strengthening financial discipline, reducing exposure to unexpected losses, and supporting more reliable managerial decisions [1]. Evidence from insurance companies has similarly indicated that effective financial risk management is associated with improved organizational performance, particularly in environments characterized by volatility and uncertainty [2]. The role of risk management is also intertwined with corporate governance, because governance mechanisms can influence the quality of risk oversight, managerial accountability, investment decisions, and the market performance of insurance firms [3]. From this perspective, understanding policyholders' risk-related behavior is as important as evaluating insurers' internal risk-management systems, since customer decisions ultimately shape premium income, portfolio composition, retention, cross-selling opportunities, and the sustainability of life insurance operations.

Life insurance differs from many other insurance products because it combines long-term risk protection with savings, investment, and intertemporal financial planning. Endowment life insurance, in particular, provides insurance coverage while accumulating a contractual value over time. Purchasing such a policy therefore reflects more than a response to a single immediate hazard; it may indicate a broader preference for future financial security, disciplined saving, protection of dependents, and reduction of uncertainty concerning long-term life events. The purchase of a second endowment life insurance policy may represent an even stronger behavioral signal because the policyholder has already experienced the first product, evaluated its costs and perceived benefits, and nevertheless decided to obtain additional coverage. In behavioral terms, repeated purchase may be associated with risk aversion, trust in insurance mechanisms, increased financial responsibility, or a desire to diversify future protection. However, such decisions are unlikely to be explained by a single variable. Age, gender, marital status, geographical location, income, education, occupation, prior insurance experience, access to insurance services, and perceptions of financial vulnerability may interact in complex ways. Research on automobile insurance decisions, for example, has demonstrated that insurance coverage choices can be shaped by sensation seeking, perceived risk, and gender-related differences, illustrating that insurance demand is influenced by both observable demographic characteristics and less directly measurable behavioral tendencies [4]. Similar behavioral complexity can be expected in the life insurance market, where decisions involve long time horizons, uncertainty about future needs, and trade-offs between present consumption and future security.

The growing complexity of insurance markets has increased the importance of advanced analytical approaches capable of identifying nonlinear patterns in customer behavior. Traditional statistical models remain valuable for hypothesis testing and estimating interpretable relationships, but they may be limited when predictors are numerous, correlated, categorical, high-dimensional, or connected through nonlinear interactions. Machine-learning algorithms can complement conventional methods by detecting complex predictive structures without requiring all relationships to be specified in advance. This capability is increasingly relevant as insurers accumulate extensive administrative data through policy-management systems, customer relationship management platforms, digital distribution channels, claims databases, and online interactions. Artificial intelligence has already been associated with improvements in insurance cost efficiency because it can automate tasks, support risk classification, and improve the allocation of organizational resources [5]. Big-data analytics and artificial intelligence have also been applied to insurance risk management and accounting transparency, demonstrating their potential to improve the processing of large financial datasets and strengthen evidence-based decision-making [6]. The development of intelligent risk-management systems therefore creates opportunities not only for underwriting and claims management but also for predicting customer acquisition, retention, lapse behavior, and demand for additional insurance products.

The broader transformation of insurance through digital technologies is commonly discussed under the concept of InsurTech. Although the term has been used in different ways, it generally refers to the application of innovative digital technologies, data infrastructures, platforms, automation tools, and new business models to insurance activities. A scientific examination of the concept has emphasized that InsurTech is not limited to isolated technological tools but represents a wider reconfiguration of insurance products, processes, institutions, and relationships with customers [7]. Contemporary research has similarly described digital technologies as increasingly embedded across insurance value chains, including distribution, underwriting, pricing, claims processing, risk prevention, and customer service [8]. The expansion of artificial intelligence, automation, and intelligent risk management across life, health, automobile, property, and specialized coverage illustrates the extent to which digital insurance solutions are reshaping both operational processes and customer experiences [9]. However, technological transformation also generates new risks. The effect of technology-related risks on insurance company performance may depend on the firm's ability to implement InsurTech innovation effectively, integrate new systems with existing operations, and manage data, cybersecurity, and organizational adaptation [10]. Therefore, the value of machine learning in insurance cannot be evaluated solely by algorithmic accuracy; it must also be considered in relation to data quality, infrastructure reliability, governance, transparency, and managerial use.

Administrative insurance data provide an especially valuable foundation for machine-learning research because they record actual behavior rather than stated intentions. Questionnaire-based studies of insurance demand and risk aversion may be affected by recall error, social desirability, hypothetical bias, and differences between what respondents say they would do and what they actually do. By contrast, transaction and policy records document observable purchasing, payment, borrowing, withdrawal, and cancellation behavior. The use of administrative data has already demonstrated predictive value in insurance-related contexts. For example, insurance claims data have been used to predict levels of long-term care need among older adults, showing that routinely collected records can support accurate classification and practical decision-making [11]. The same principle can be extended to life insurance: records available at the time of the first policy purchase may contain signals that help predict whether a customer will later purchase additional coverage. Nevertheless, administrative databases also present methodological challenges. They may contain incomplete fields, inconsistent coding, duplicate records, nonstandard categories, and values generated by different operational systems. Accounting information systems and internal controls are therefore important because the quality and reliability of organizational data affect employee performance, reporting processes, and managerial decisions [12]. In machine-learning applications, inadequate data governance can lead to biased estimates, unstable predictions, and misleading managerial conclusions.

The increasing reliance on digital systems has also broadened the risk landscape faced by insurance companies and their customers. Cyber insurance has emerged as a mechanism for mitigating digital and international trade risks, indicating that insurers are increasingly required to assess threats that are technologically complex and difficult to quantify [13]. Legal research has similarly emphasized the role of cyber insurance in reducing legal risks arising from digital trade and technology-based transactions [14]. At the corporate level, social responsibility may function as a form of reputational or organizational insurance against cybersecurity incidents, suggesting that risk protection can involve not only formal insurance contracts but also broader organizational resources and stakeholder relationships [15]. These developments are relevant to life insurance analytics because they illustrate a wider shift from static, product-centered risk assessment toward dynamic, data-driven, and multidimensional risk management. As insurers adopt more sophisticated information technologies, they gain opportunities to understand customers more precisely, but they also assume greater responsibilities regarding privacy, model governance, cybersecurity, and the ethical use of personal information.

Insurance-company competitiveness increasingly depends on the ability to translate data into actionable knowledge. Financial insurers operate in markets shaped by customer expectations, technological change, regulatory requirements, capital constraints, and competition from digital platforms. Strategies for enhancing competitiveness therefore require improvements in service quality, product differentiation, digital capabilities, customer segmentation, and the efficiency of risk assessment [16]. Artificial intelligence can contribute to these strategies by enabling insurers to identify high-potential customers, personalize communication, optimize marketing resources, and design products suited to particular demographic or regional groups. However, prediction must be accompanied by sound actuarial and financial judgment. Actuaries continue to play a fundamental role in assessing financial risks, estimating liabilities, evaluating uncertainty, and supporting the solvency of insurance companies [17]. Machine-learning systems should therefore be viewed as complementary analytical instruments rather than substitutes for actuarial expertise. The strongest decision-making framework is likely to combine algorithmic pattern recognition with actuarial reasoning, managerial interpretation, regulatory oversight, and knowledge of insurance-market behavior.

The financial management of insurance companies further demonstrates the necessity of integrating customer-level prediction with portfolio-level risk control. Insurers invest accumulated premiums and reserves, and their long-term ability to honor obligations depends on effective asset allocation and risk measurement. Portfolio optimization using copula functions and value-at-risk approaches has shown how insurance companies can improve the assessment of dependencies and downside risk in investment portfolios [18]. Risk-transfer mechanisms also extend beyond primary insurance to reinsurance arrangements. Stop-loss reinsurance thresholds, particularly under dependent risks, require careful mathematical assessment because loss dependencies can significantly influence the insurer’s retained exposure [19]. In banking and related financial sectors, market development, financial crises, and deposit insurance have been shown to influence institutional risk-taking, further demonstrating that insurance arrangements can alter behavior and incentives within financial systems [20]. These findings reinforce the broader principle that insurance decisions are embedded within systems of incentives, expectations, and risk perceptions. At the consumer level, the availability of life insurance may reduce perceived uncertainty, while the decision to purchase multiple policies may indicate a particularly strong preference for transferring or managing long-term financial risk.

The interpretation of repeated life insurance purchases must also consider organizational and environmental uncertainty. Severe crises can expose weaknesses in conventional enterprise risk-management systems, making organizational resilience essential for maintaining operations and responding to novel threats [21]. For insurance companies, resilience includes the ability to adapt underwriting models, maintain customer services, process claims, preserve liquidity, and revise products in response to economic shocks. Climate-related risks have similarly encouraged the development of digital insurance mechanisms capable of improving monitoring, pricing, and responses to climate fluctuations [22]. At the same time, the sustainability of the insurance industry depends on its capacity to remain financially viable, socially relevant, and responsive to emerging environmental and technological risks [23]. Life insurance development contributes to this sustainability by expanding long-term protection and mobilizing savings, yet its growth requires a precise understanding of why some customers purchase additional coverage while others do not.

Risk aversion is not directly observable in administrative databases, but it can be inferred through patterns of actual financial behavior when the conceptual and operational definitions are specified carefully. Individuals who purchase an additional endowment life insurance policy despite having alternative uses for their funds may display a stronger preference for contractual protection, future certainty, and long-term accumulation. Nevertheless, repeated purchase can also be influenced by marketing exposure, institutional affiliation, access to insurance representatives, previous satisfaction, regional culture, family responsibilities, and prior ownership of other insurance products. For this reason, predictive modeling is particularly suitable: it does not assume that all customers follow an identical decision process and can identify combinations of variables that distinguish second-policy purchasers from nonpurchasers. XGBoost is appropriate in this setting because it can model nonlinear relationships, accommodate interactions, regulate tree complexity, process large datasets efficiently, and generate feature-importance measures. These characteristics make it possible to move beyond a purely descriptive analysis and produce a model with direct managerial relevance.

The managerial value of such a model lies in its capacity to support targeted rather than indiscriminate marketing. Insurance companies often allocate substantial resources to customer contact, sales campaigns, agent incentives, and product promotion. When these activities are directed toward customers without considering purchase propensity, conversion rates may remain low and marketing costs may increase. A predictive model can rank customers according to their estimated probability of purchasing a second policy, enabling customer relationship management units to prioritize high-potential cases. It can also reveal whether certain age groups, geographical areas, marital-status categories, or holders of other insurance policies are more likely to purchase additional life coverage. This information can support market segmentation, regional planning, product customization, and the design of differentiated communication strategies. The value of performance-linked incentives in reducing promotion-related uncertainty has been examined in organizational finance, illustrating how contractual and informational mechanisms can function as forms of protection against uncertain outcomes [24]. In a comparable manner, predictive customer analytics can reduce uncertainty in insurance marketing by improving the allocation of limited sales resources.

Despite the rapid development of artificial intelligence and digital insurance, empirical research on the prediction of second endowment life insurance purchases remains limited, particularly when based on large-scale real-world data from the Iranian insurance market. Many previous studies have focused on organizational risk management, financial performance, underwriting, cyber risk, investment portfolios, or broad digital transformation. Fewer studies have examined repeated life insurance purchasing as a behavioral indicator of risk aversion using operational policy records. Moreover, the Iranian market presents distinctive conditions, including economic instability, inflationary pressures, heterogeneous regional insurance cultures, organizational group-insurance arrangements, and varying levels of access to life insurance services. These conditions may influence both the perceived attractiveness of endowment life insurance and the observable characteristics of customers who purchase additional coverage. An analysis based on actual policy records can therefore contribute to the literature by connecting risk-aversion theory, customer behavior, InsurTech, and machine-learning prediction within a specific emerging-market context.

Accordingly, the aim of this study was to develop and validate an XGBoost-based machine-learning model for identifying the demographic, geographical, behavioral, and insurance-related factors associated with policyholders' risk aversion and predicting the probability of purchasing a second endowment life insurance policy.

2. Methodology

This study adopted an applied, retrospective, predictive research design based on administrative insurance data and supervised machine-learning methods. Its primary objective was to develop and evaluate an Extreme Gradient Boosting model for predicting the probability that a policyholder would purchase a second endowment life insurance policy. The purchase of an additional life insurance policy was treated as an observable behavioral indicator of risk aversion, particularly because policyholders made this decision under unstable economic conditions in which alternative investment opportunities were available. Rather than measuring risk aversion through self-report questionnaires, the study examined actual insurance-purchasing behavior recorded in operational information systems. Accordingly, the research design combined behavioral risk analysis, customer analytics, and predictive modeling to identify the demographic, socioeconomic, and insurance-related characteristics associated with the purchase of a second endowment life insurance policy.

The study population consisted of individuals who had purchased an initial endowment life insurance policy from Sarmad Insurance Company and whose policyholder and insured-person identities were the same. The observation period extended from March 2021 to March 2024. The records were obtained from Sarmad Insurance Company, one of the major insurance companies operating in Iran, and Fanavaran Company, which provides an insurance technology platform used for policy issuance and the recording of insurance transactions. All eligible records available within the defined observation period were included. Therefore, conventional probability sampling was not performed, and a census approach was employed. The use of all eligible records increased the statistical and computational power of the predictive model, reduced sampling-related estimation error, and improved the model's ability to learn relatively uncommon patterns associated with the purchase of a second life insurance policy.

The initial databases contained policy records with different levels of completeness, consistency, and structural quality. In particular, records associated with second life insurance policies had been generated through relatively new software systems and were sometimes stored in nonstandard or incomplete formats. Consequently, the raw data showed substantial fragmentation, limited informational depth, duplicated entries, implausible values, and inconsistencies across the two data sources. A rigorous eligibility assessment and data-cleaning procedure was therefore implemented before the construction of the analytical dataset. Only policies that had not been terminated by the insurer before the end of the first policy year for risk-management reasons were retained. This criterion was applied to exclude contracts that did not represent stable or genuine long-term participation in endowment life insurance.

Two additional exclusion criteria were used to reduce behavioral misclassification and ensure that the retained observations appropriately represented policyholders' insurance and risk-related decisions. First, policies that were terminated or involved a withdrawal from the accumulated policy value during the first year were excluded. Early termination and first-year withdrawals may be motivated by reasons unrelated to long-term insurance demand, including the temporary use of insurance policies for tax-related purposes. Such records could therefore distort the interpretation of risk aversion by classifying strategically or temporarily purchased policies as genuine long-term insurance decisions. Second, policyholders who borrowed against their accumulated policy value from the second year onward and subsequently failed to repay the loan installments were excluded. Although borrowing against up to approximately 90% of the accumulated value is a legitimate policy benefit, failure to repay may cause the contract to continue only until the remaining accumulated value is exhausted. This behavior does not necessarily indicate a stable preference for risk protection and was therefore considered unsuitable for predicting risk-averse insurance-purchasing behavior.

After the eligibility criteria and preprocessing procedures had been applied, the final analytical dataset comprised 114,566 policyholders. Of these, 18,604 had purchased a second endowment life insurance policy and were assigned to the positive outcome class, whereas 95,962 had purchased only one endowment life insurance policy and had not completed a second purchase by the end of the observation period. The prediction horizon extended from the date of the first life insurance purchase to the end of the available observation period. Thus, the model did not predict second purchases within a fixed interval such as six months or one year. Instead, it estimated whether a policyholder would purchase a second policy at any point during the remaining period covered by the administrative database.

The dependent variable was defined as a binary outcome. Policyholders who purchased a second endowment life insurance policy after their initial purchase were coded as 1, whereas those who retained only their first policy and made no second purchase before the end of the observation period were coded as 0. To prevent data leakage, only information available to the insurance company at the time of the initial policy purchase was used as model input. Variables created after the second purchase, affected by the second purchase, or indirectly revealing the outcome were excluded. This restriction ensured that the model's performance reflected its genuine ability to predict future customer behavior rather than its ability to identify information generated after the outcome had already occurred.

All records were anonymized before analysis. Names, identification numbers, contact information, and other directly identifiable information were removed or replaced with nonidentifying codes. The data were used exclusively for scientific purposes in accordance with research ethics principles, organizational data-protection requirements, and the internal regulations of Sarmad Insurance Company.

The required information was collected through documentary analysis of secondary administrative data rather than through questionnaires, interviews, or experimental measurements. The principal data-collection tools were the operational database of Sarmad Insurance Company and the information system maintained by Fanavaran Company. The Sarmad Insurance database was used to identify policyholders who had purchased first and second endowment life insurance policies, whereas the Fanavaran platform was used to retrieve insurance histories and other available policyholder characteristics. The records reflected actual customer transactions and insurance histories documented during routine organizational operations. Consequently, the dataset was less vulnerable to recall bias, social desirability bias, and common-method bias than data obtained exclusively through self-report instruments.

The extraction and preparation procedure was conducted in several sequential stages. Initially, the target population, outcome variable, observation period, and required customer characteristics were defined in accordance with the research objectives. Relevant records were then extracted from the two information systems and integrated by means of shared nonidentifying record fields. The integrated database was subsequently examined for duplicate observations, missing values, inconsistent codes, unrealistic numerical values, and discrepancies between corresponding fields in the two systems. Duplicate records were removed, implausible values were investigated, and inconsistent categories were reconciled wherever reliable source information was available. Missing values were assessed according to their frequency, pattern, and relevance to the predictive variables. Records or fields with inadequate informational quality were excluded or treated through appropriate preprocessing procedures, depending on the extent and nature of the missingness.

The predictors were selected according to the theoretical foundations of insurance demand, the available administrative information, and their temporal availability at the time of the first policy purchase. Thirteen demographic, behavioral, socioeconomic, and insurance-related variables were initially entered as candidate predictors. These included age, gender, marital status, city of residence, province of residence, occupation, income, educational attainment, possession of other insurance policies, possession of automobile insurance, and other available customer and insurance-history characteristics. Automobile insurance included third-party liability coverage and comprehensive motor insurance. The primary analytical findings subsequently indicated that age, gender, city and province of residence, marital status, possession of other insurance policies, and possession of automobile insurance were among the most influential predictors of purchasing an additional endowment life insurance policy. Variables such as occupation, income, and educational attainment demonstrated relatively limited predictive importance, partly because a large proportion of the records had been obtained from employees of major organizational clients and therefore showed limited variability in these characteristics.

Before modeling, all variables were transformed into formats compatible with machine-learning analysis. Numerical fields were checked for impossible ranges and extreme observations, while categorical variables were standardized to eliminate inconsistent spelling, coding, and category definitions. Where required, categorical predictors were encoded numerically. The resulting structured dataset represented one of the principal methodological contributions of the study because it was developed from large-scale real-world insurance records without reliance on international databases or externally prepared datasets.

Data processing, model development, training, validation, and evaluation were conducted in the Python programming environment. The predictive model was implemented specifically for this study using the XGBoost algorithm. XGBoost was selected because the outcome was binary, the dataset was large and heterogeneous, and the relationships between customer characteristics and insurance-purchasing behavior were expected to be nonlinear and interactive. The algorithm constructs an ensemble of sequential decision trees, with each new tree attempting to correct the prediction errors produced by the preceding trees. It is also capable of incorporating regularization, controlling tree complexity, and calculating feature importance, making it appropriate for both prediction and the identification of factors associated with second-policy purchases.

Following preprocessing, the dataset was randomly divided into a training set containing 80% of the observations and a test set containing the remaining 20%. The training set was used to estimate the model and learn patterns distinguishing policyholders who purchased a second policy from those who did not. The test set was withheld during model development and used only to evaluate the final model on previously unseen observations. This separation was implemented to assess generalizability and reduce the likelihood that model performance would be overstated because of overfitting.

The outcome distribution was substantially imbalanced because the number of policyholders who did not purchase a second policy was considerably larger than the number who did. The positive class comprised 18,604 policyholders, while the negative class comprised 95,962 policyholders. To reduce bias toward the majority class and improve the identification of second-policy purchasers, the XGBoost scale_pos_weight parameter was calculated as the ratio of negative to positive observations. The resulting value was approximately 5.15, calculated by dividing 95,962 by 18,604. This weighting increased the penalty associated with incorrectly classifying members of the minority class and allowed the algorithm to assign greater importance to the accurate identification of policyholders who purchased a second policy.

Hyperparameter tuning was performed before the final model was trained. The parameters examined included the number of decision trees, tree depth, learning rate, the proportion of observations sampled for each tree, the proportion of predictors sampled for each tree, the minimum loss reduction required for an additional split, the minimum child weight, and L1 and L2 regularization coefficients. The final model used 200 estimators, a learning rate of 0.05, a maximum tree depth of 4, a minimum child weight of 3, a gamma value of 0.10, a subsample ratio of 0.85, and a column-sampling ratio of 0.80. The L1 regularization coefficient was set at 0.10, and the L2 regularization coefficient was set at 1. These settings were selected according to cross-validation performance and were intended to balance predictive accuracy, computational efficiency, and resistance to overfitting.

K-fold cross-validation was applied to the training data to evaluate the stability of the model and reduce dependence on a single training-validation partition. In this procedure, the training dataset was divided into approximately equal subsets. During each iteration, one subset served as the validation set and the remaining subsets were used for model training. The process was repeated until each subset had served once as the validation set, and the results were averaged to obtain a more robust estimate of predictive performance. Hyperparameter selection was based on the model's aggregated cross-validation results rather than its performance on one arbitrary partition.

The final model was evaluated on the independent test set using the confusion matrix, accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve. Accuracy represented the proportion of all observations classified correctly. Precision indicated the proportion of customers predicted to purchase a second policy who actually did so, whereas recall measured the proportion of actual second-policy purchasers correctly identified by the model. The F1 score provided a harmonic balance between precision and recall and was particularly relevant because of the unequal class distribution. The area under the receiver operating characteristic curve assessed the model's ability to discriminate between purchasers and nonpurchasers across different classification thresholds. Feature-importance measures generated by XGBoost were also examined to determine the relative contribution of the candidate predictors to the model's decisions. The simultaneous use of discrimination, classification, and feature-importance measures enabled a comprehensive evaluation of both the predictive performance and practical interpretability of the model.

3. Findings and Results

The data used in the present study were extracted from the administrative database of Sarmad Insurance Company and its insurance technology platform operated by Fanavaran Company. Following data screening, removal of outliers, correction of inconsistent values, elimination of duplicate observations, and preparation of the variables for machine-learning analysis, 95,962 valid records were retained from the available administrative data. The dataset contained information on individuals who had purchased an initial endowment life insurance policy, among whom 18,604 subsequently purchased a second endowment life insurance policy. Accordingly, the dependent variable was binary and represented the occurrence or nonoccurrence of a second life insurance purchase. Policyholders who purchased an additional endowment life insurance policy were coded as 1, whereas those who did not make a second purchase during the observation period were coded as 0.

The descriptive examination covered both quantitative and qualitative variables. For quantitative variables, the number of valid observations, mean, median, standard deviation, minimum, maximum, and range were examined to assess the central tendency and dispersion of the data. For qualitative variables, the frequencies and percentages of the categories were reviewed to describe the demographic, geographical, socioeconomic, and insurance-related composition of the study population. The distribution of the outcome variable showed a substantial class imbalance because the number of policyholders who purchased a second life insurance policy was considerably lower than the number of policyholders who did not. Therefore, after hyperparameter tuning, the scale_pos_weight parameter was incorporated into the model to reduce majority-class bias and strengthen the model's ability to identify customers belonging to the minority class of second-policy purchasers.

The XGBoost model was trained using the cleaned and prepared administrative data. Before model training, categorical predictors were transformed into numerical representations suitable for machine-learning analysis. Gender was encoded as a binary variable. Marital status was transformed using one-hot encoding, thereby generating separate indicator variables for the relevant marital-status categories. Province and city were converted into numerical values through target encoding because these variables contained numerous geographical categories, and conventional one-hot encoding would have generated an excessively high-dimensional feature space. Policyholders' ages were calculated from their recorded years of birth and included as numerical input variables.

Following feature engineering and variable encoding, the final input matrix contained 97 predictors. The increase from the original set of variables to 97 model features resulted primarily from the transformation of categorical variables into multiple numerical indicators and encoded geographical features. The resulting dataset was divided into training and test subsets. Eighty percent of the observations were allocated to the training set, while the remaining 20% were reserved for independent model evaluation. A stratified train–test split was used to preserve the relative distribution of second-policy purchasers and nonpurchasers in both subsets. This procedure was particularly important because a simple random split could have produced unequal class distributions and consequently distorted the evaluation of model performance.

Table 1. Preparation and Configuration of the XGBoost Model

Characteristic	Value
Algorithm	XGBoost
Type of problem	Binary classification
Number of records used for modeling	95,962
Number of final input features	97
Training-set proportion	80%
Test-set proportion	20%
Data-partitioning method	Stratified train–test split
Validation method	10-fold cross-validation
Target variable	Purchase of a second endowment life insurance policy

As shown in Table 1, the XGBoost algorithm was implemented as a binary classification model using 95,962 records and 97 final features. The stratified partitioning procedure maintained the distribution of the purchase and nonpurchase classes in the training and test sets, thereby permitting a more reliable evaluation. In addition, k-fold cross-validation was used so that the model's performance would not depend exclusively on a single partition of the data.

Cross-validation was conducted to determine whether the model could maintain its predictive performance when applied to observations that had not been used during a particular training iteration. A model that performs well only on its training data but shows a considerable decline when exposed to new data is likely to be affected by overfitting and may have limited practical generalizability. The use of k-fold cross-validation reduces dependence on a single data split and produces a more stable estimate of out-of-sample performance.

Considering the large volume of data, the binary classification structure of the outcome, the imbalance between the two classes, and the need to develop a stable predictive model, 10-fold cross-validation was selected. The dataset used in the cross-validation procedure was divided into 10 approximately equal folds. In each iteration, nine folds were used to train the model and the remaining fold was used for validation. This process was repeated 10 times, with each fold serving exactly once as the validation subset. The average values obtained across the 10 iterations were then used as estimates of the model's overall validation performance.

Table 2. Ten-Fold Cross-Validation Performance of the XGBoost Model

Fold	Accuracy	Precision	Recall	F1 Score	ROC–AUC
1	0.748	0.374	0.761	0.502	0.864
2	0.754	0.382	0.772	0.510	0.871
3	0.758	0.380	0.776	0.512	0.873
4	0.760	0.384	0.781	0.515	0.876
5	0.756	0.379	0.775	0.509	0.869
6	0.753	0.376	0.771	0.506	0.868
7	0.757	0.381	0.778	0.512	0.874
8	0.755	0.377	0.773	0.507	0.871
9	0.759	0.383	0.779	0.514	0.875
10	0.756	0.378	0.774	0.508	0.870
Mean	0.756	0.379	0.774	0.510	0.870

The cross-validation results showed that the XGBoost model maintained a stable level of performance across all 10 folds. Accuracy ranged from 0.748 to 0.760, while ROC–AUC ranged from 0.864 to 0.876. The mean accuracy was 0.756, indicating that approximately 75.6% of observations were correctly classified across the validation iterations. The mean ROC–AUC was 0.870, demonstrating that the model had a strong ability to discriminate between customers who purchased a second life insurance policy and those who did not.

The mean recall was 0.774, indicating that the model correctly identified approximately 77.4% of the actual second-policy purchasers. The mean precision was lower, at 0.379, showing that a proportion of customers predicted to be second-policy purchasers did not actually complete a second purchase during the observation period. The mean F1 score was 0.510, representing the harmonic balance between precision and recall. The close proximity of the performance indicators across the 10 folds demonstrated that the model had limited sensitivity to a particular division of the dataset. This consistency supports the stability and generalizability of the model for predicting repeated life insurance purchasing behavior.

Following training and cross-validation, the final XGBoost model was evaluated on the independent test dataset using standard binary-classification measures. Because the principal objective was to identify customers likely to purchase a second endowment life insurance policy, particular attention was given to the performance of the model for the positive class. Accuracy, precision, recall, F1 score, ROC–AUC, and the confusion matrix were considered simultaneously because each measure represents a different aspect of classification performance.

Accuracy indicates the percentage of all customers whose purchasing status was correctly classified. Precision indicates the proportion of customers classified as likely second-policy purchasers who actually completed a second purchase. Recall represents the proportion of all actual second-policy purchasers successfully identified by the model. The F1 score combines precision and recall into a single measure and is especially useful when the classes are imbalanced. ROC–AUC evaluates the ability of the model to rank positive cases above negative cases across all possible classification thresholds.

Table 3. Performance Indicators of the Final XGBoost Model

Performance indicator	Value
Accuracy	75.56%
ROC–AUC	87.05%
Precision	37.83%
Recall	77.41%
F1 score	50.82%

As presented in Table 3, the final XGBoost model achieved an accuracy of 75.56%, meaning that approximately three quarters of the observations in the test dataset were correctly classified. The ROC–AUC value was 87.05%, indicating a strong discriminatory capacity. This result shows that the model was effective in ranking customers according to their probability of purchasing an additional life insurance policy.

The recall value of 77.41% was particularly important for the practical objective of the study. It indicated that the model successfully detected more than three quarters of customers who actually purchased a second endowment life insurance policy. Precision was 37.83%, suggesting that the model also classified a considerable number of nonpurchasers as potential second-policy purchasers. The F1 score calculated from the reported precision and recall values was 50.82%. Although precision was moderate, the relatively high recall reflected the model configuration's emphasis on identifying as many potential purchasers as possible.

The confusion matrix was examined to provide a more detailed interpretation of correct and incorrect classifications. Unlike overall accuracy, which summarizes all correct predictions in a single value, the confusion matrix distinguishes true negatives, true positives, false positives, and false negatives. The model correctly classified 14,366 customers who did not purchase a second life insurance policy as nonpurchasers. These observations represented true negatives and constituted the largest category of correct predictions.

The model also correctly identified 2,882 customers who purchased a second life insurance policy. These observations represented true positives and demonstrated the model's capacity to identify a substantial proportion of customers belonging to the minority class. A total of 4,737 customers who did not purchase a second policy were nevertheless classified as likely purchasers; these cases represented false positives. In addition, 841 customers who actually purchased a second policy were incorrectly classified as nonpurchasers and therefore represented false negatives.

The confusion-matrix values were consistent with the reported recall of 77.41%, because the model correctly identified 2,882 of the 3,723 actual second-policy purchasers in the test dataset. The relatively small number of false negatives was especially relevant to the research objective. False negatives represent customers who genuinely purchased a second policy but were not recognized by the model as potential purchasers. In a targeted insurance-marketing context, failure to identify such customers may result in the loss of valuable sales opportunities.

By contrast, false positives represent customers who were classified as potential second-policy purchasers but did not purchase an additional policy during the study period. A relatively larger number of false positives is expected in an imbalanced classification problem when the model is deliberately configured to increase detection of the minority class. From a managerial perspective, contacting a customer who ultimately does not purchase an additional policy generally entails a lower cost than failing to identify a customer with a genuine propensity to make a second purchase. Therefore, the observed balance between false positives and false negatives may be considered appropriate for customer targeting, campaign design, and sales-lead prioritization.

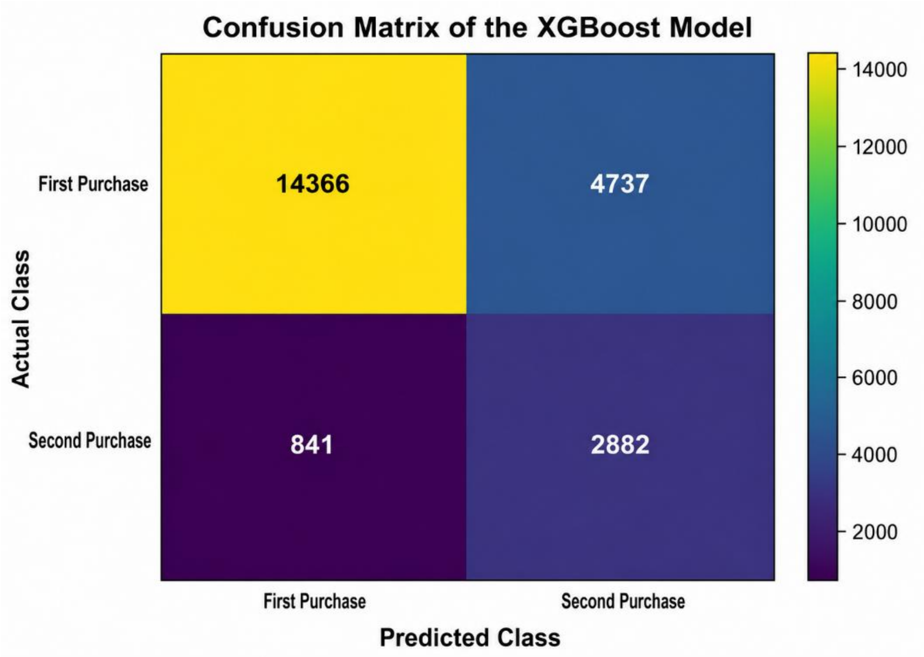


Figure 1. Confusion Matrix of the XGBoost Model for Predicting the Purchase of a Second Endowment Life Insurance Policy

Overall, the confusion-matrix analysis showed that the model achieved an acceptable balance between correctly identifying second-policy purchasers and controlling classification errors. Its relatively high recall suggests that the algorithm can support insurance companies in identifying prospective customers, designing targeted marketing campaigns, allocating customer relationship management resources, improving conversion rates, and reducing the likelihood that high-potential customers will be overlooked.

The feature-importance analysis was conducted to identify the predictors that contributed most strongly to the model's classification decisions. The results showed that demographic, geographical, and behavioral characteristics, including previous purchases of other forms of insurance coverage, had the greatest influence on the prediction of a second endowment life insurance purchase. The variables did not contribute equally to the model. Instead, XGBoost distinguished and ranked the predictors according to the frequency and usefulness of their participation in decision-tree splits.

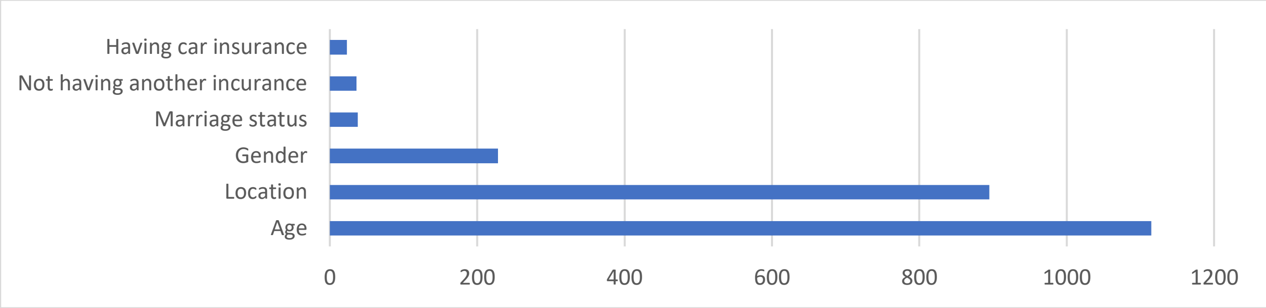


Figure 2. Overall Importance of Predictor Variables in the Purchase of a Second Endowment Life Insurance Policy

Age received the highest importance score, with a value of 1,154, and was therefore the most influential predictor of second-policy purchasing behavior. This finding indicates that the probability of purchasing an additional life insurance policy differed considerably across age groups. From an insurance-marketing perspective, the prominence of age is plausible because financial needs, accumulated income, family responsibilities, perceptions of vulnerability, future planning, and attitudes toward long-term financial security change throughout the life course. These age-related changes may substantially influence the perceived value of obtaining additional endowment life insurance coverage.

Place of residence ranked second, with an importance score of 895. The strong contribution of geographical variables indicates that regional characteristics played a substantial role in customer behavior. Differences among cities and provinces may reflect variations in economic development, household purchasing power, access to insurance branches and representatives, local insurance culture, employment structures, social conditions, and familiarity with long-term financial products. The importance of residence suggests that marketing strategies and insurance products should not necessarily be designed uniformly across all regions.

Gender ranked third, with an importance score of 228. Although its contribution was considerably lower than those of age and residence, it remained a meaningful predictor. This result suggests that differences in financial decision-making, perceived risk, family roles, saving behavior, and attitudes toward insurance between women and men may affect the likelihood of purchasing an additional life insurance policy.

Marital-status categories, including single, married, and divorced status, had lower importance scores. Nevertheless, their presence among the selected predictors showed that family status was not entirely unrelated to repeated insurance purchasing. Marital status may influence financial obligations, household decision-making, dependency structures, and the perceived need for future protection. However, its predictive contribution was more limited than that of age and geographical location.

Variables representing customers' previous insurance histories also appeared among the influential features. In particular, possession of other insurance policies and automobile insurance, including third-party liability and comprehensive motor insurance, contributed to the prediction of a second life insurance purchase. Their importance indicates that prior interaction with insurance companies and familiarity with insurance products may provide valuable information about future purchasing behavior. Customers who already hold several forms of insurance coverage may possess greater insurance awareness, higher trust in insurance institutions, or a stronger general preference for risk-transfer mechanisms.

Conversely, occupation, income, educational attainment, and several other variables showed relatively limited predictive importance. One explanation was the high degree of similarity among the employees of large organizational clients from whom a substantial proportion of the insurance records had been obtained. Limited variability in occupational, income, and educational characteristics reduced their capacity to differentiate second-policy purchasers from nonpurchasers.

Taken together, the feature-importance results demonstrated that age, geographical location, gender, marital status, possession of other insurance policies, and automobile insurance history were the principal factors associated with purchasing a second endowment life insurance policy. The findings also demonstrated the ability of XGBoost to identify influential variables and capture complex and potentially nonlinear relationships among customer characteristics. These results can support insurance companies in customer segmentation, market targeting, regional strategy formulation, campaign personalization, product development, and managerial decision-making concerning the expansion of endowment life insurance.

4. Discussion and Conclusion

The present study developed and evaluated an XGBoost-based machine-learning model for predicting the purchase of a second endowment life insurance policy and identifying the principal demographic, geographical, behavioral, and insurance-related factors associated with this decision. The findings showed that the model achieved an overall accuracy of 75.56%, a ROC–AUC value of 87.05%, a recall of 77.41%, a precision of 37.83%, and an F1 score of approximately 50.82%. The 10-fold cross-validation results were similarly stable, with a mean accuracy of 0.756 and a mean ROC–AUC of 0.870. These findings indicate that the model possessed strong discriminatory ability and could distinguish customers who were more likely to purchase a second endowment life insurance policy from those who were unlikely to do so. The relatively high recall was particularly important because the principal practical objective was to identify as many potential second-policy purchasers as possible. Although precision was more moderate, the model correctly identified more than three quarters of actual second-policy purchasers, which demonstrates its usefulness as a customer-screening and marketing-support instrument rather than as an entirely deterministic decision rule.

The strong ROC–AUC value supports the growing body of evidence indicating that artificial intelligence and machine-learning techniques can improve analytical efficiency in the insurance industry. Artificial intelligence enables insurers to process large administrative datasets, identify relationships that may not be evident through conventional descriptive analyses, and improve the allocation of underwriting and marketing resources [5]. The present findings are also consistent with the broader development of InsurTech, in which digital technologies are increasingly applied across underwriting, pricing, customer management, claims processing, and risk assessment [8]. InsurTech represents not only the digitization of existing insurance operations but also a transformation in how insurers define risks, interact with customers, and generate value from information [7]. Accordingly, the satisfactory predictive performance obtained in this study demonstrates how administrative life insurance records can be converted into actionable managerial knowledge through a carefully developed machine-learning model.

The findings further align with research emphasizing the role of artificial intelligence and big-data analytics in improving risk management and accounting transparency in insurance companies. The use of large-scale data can strengthen the accuracy of managerial evaluations, facilitate the identification of risk patterns, and improve organizational responsiveness [6]. Similarly, digital insurance solutions based on artificial intelligence, automation, and intelligent risk management have been described as increasingly relevant across life, health, automobile, property, and specialized insurance products [9]. The current study extends these applications to repeated endowment life insurance purchasing by demonstrating that customer information available at the time of the first policy purchase can be used to estimate the likelihood of a subsequent purchase. This is an important contribution because much of the existing literature has concentrated on underwriting risk, claims, fraud, financial performance, or insurer-level risk management rather than on predicting repeated life insurance purchases as an indicator of customers' preference for financial protection.

The stability of the model across the 10 validation folds is another noteworthy finding. The limited variation in accuracy, recall, precision, F1 score, and ROC–AUC across the folds suggests that the model was not excessively dependent on one particular division of the data. This result reduces concerns regarding overfitting and supports the potential generalizability of the model to new customer records drawn from a similar market environment. The use of large administrative datasets has previously proved valuable in insurance-related prediction. For example, insurance claims data have been successfully used to predict long-term care needs among older adults, indicating that routinely collected insurance records may contain sufficient information for practically meaningful classification [11]. The present results support the same methodological principle in life insurance: administrative records created through actual transactions can provide a more behaviorally grounded basis for prediction than hypothetical responses obtained through self-report questionnaires.

The confusion-matrix findings showed that the model correctly classified 14,366 customers who did not purchase a second policy and 2,882 customers who did. It incorrectly classified 4,737 nonpurchasers as likely purchasers and failed to identify 841 actual second-policy purchasers. This distribution explains the combination of relatively high recall and moderate precision. The model was calibrated to identify a large proportion of the minority class, and this increased the number of false-positive predictions. In the context of insurance marketing, however, this trade-off may be acceptable. Contacting some customers who do not ultimately purchase a second policy is generally less costly than failing to identify customers who have a high propensity to purchase. From a risk-management perspective, model evaluation should therefore be connected to the economic consequences of classification errors rather than based solely on overall accuracy. Insurance risk management is intended to support organizational performance by improving the quality of decisions under uncertainty [1]. Financial risk-management practices can likewise strengthen organizational performance when analytical information is used to direct resources toward strategically relevant risks and opportunities [2]. Thus, the model's relatively high sensitivity to potential purchasers may offer greater managerial value than a model with higher precision but substantially lower recall.

The variable-importance analysis showed that age was the most influential predictor of purchasing a second endowment life insurance policy. This finding indicates that repeated life insurance purchasing varies substantially across the life course. Age may affect individuals' perceptions of mortality and financial vulnerability, their accumulated savings, their responsibilities toward spouses or children, their retirement planning, and their preference for long-term security. Younger policyholders may have different priorities from middle-aged customers with growing family obligations, while older customers may display greater concern about retirement income, estate planning, or the financial protection of dependents. The importance of age therefore supports the interpretation of second-policy purchasing as a multidimensional decision involving both risk aversion and changing financial needs.

The result is also conceptually consistent with evidence showing that insurance decisions are shaped by behavioral and perceptual characteristics rather than by purely financial calculations. Research on automobile insurance has demonstrated that risk perception, gender, and sensation-seeking tendencies influence decisions concerning insurance coverage [4]. Although automobile and life insurance address different types of risk, both involve individual judgments about uncertainty, expected loss, and the value of transferring risk to an insurer. Age-related differences in risk perception may therefore partly explain why age emerged as the strongest predictor in the present study. Moreover, life insurance is a long-term financial product, and its perceived value is likely to change as customers progress through different stages of employment, marriage, parenthood, asset accumulation, and retirement preparation.

Geographical location was the second most influential predictor. The importance of city and province suggests that repeated insurance purchasing is embedded in regional economic, institutional, and cultural conditions. Differences in household income, employment stability, access to insurance branches, density of sales representatives, financial literacy, trust in insurance companies, and local insurance culture may influence the probability of purchasing additional coverage. Regional economic instability may also affect how customers allocate financial resources between insurance, property, foreign currency, gold, savings accounts, and other investment options. The relevance of location is compatible with the view that insurers must develop differentiated competitive strategies instead of assuming homogeneous market behavior across all regions. Improving competitiveness requires insurers to understand customer segments, respond to local market conditions, and align products and distribution mechanisms with the characteristics of target populations [16].

The geographical result may also be interpreted in light of the increasing importance of digital insurance. Digital channels can reduce geographical barriers by allowing customers to obtain information, compare products, communicate with insurers, and purchase coverage remotely. Digital insurance has been identified as a potential mechanism for improving access to risk-protection products and responding to emerging risks, including those associated with environmental and climate fluctuations [22]. Nevertheless, technology adoption remains uneven across regions, and the performance benefits of InsurTech depend on insurers' capacity to manage technological risks, infrastructure limitations, and organizational adaptation [10]. Therefore, the effect of city and province may reflect not only regional economic and cultural differences but also unequal access to digital platforms and insurance services.

Gender ranked as the third most influential variable. This finding suggests that men and women may differ in their perceptions of future financial risk, family obligations, saving behavior, confidence in insurance, and willingness to purchase additional long-term coverage. The result is aligned with evidence that gender can moderate the relationship between psychological tendencies, risk perception, and automobile insurance decisions [4]. Gender differences in life insurance purchasing should not be interpreted as fixed or universal, because they may be shaped by labor-market participation, household decision-making roles, income control, financial literacy, and cultural norms. Nonetheless, the result indicates that gender contributes meaningful predictive information when combined with other customer characteristics.

Marital status had a lower but still relevant predictive contribution. Married customers may have greater responsibility for dependents and may therefore perceive stronger benefits from additional life insurance. Single or divorced individuals may have different financial priorities, levels of dependency, or long-term protection needs. However, the lower importance of marital status relative to age and residence indicates that family status alone is insufficient to explain repeated purchasing. Its effect may depend on interactions with age, income, gender, number of dependents, employment, and existing insurance coverage. The ability of XGBoost to model such nonlinear interactions is one reason for its suitability in the present study. Conventional models may estimate an average independent effect for marital status, whereas a tree-based ensemble can capture situations in which marital status becomes important only within particular demographic or regional subgroups.

Possession of other insurance policies and automobile insurance also contributed to the prediction of a second endowment life insurance purchase. This finding indicates that previous insurance experience may be an important behavioral marker of future demand. Customers who already hold automobile, third-party liability, comprehensive motor, or other insurance policies may be more familiar with insurance concepts, more trusting of insurance institutions, and more willing to use formal risk-transfer mechanisms. Prior insurance ownership may also indicate a broader orientation toward financial protection and compliance with risk-management practices. This interpretation is consistent with the principle that insurance decisions are interconnected rather than isolated. Deposit insurance, for instance, can influence institutional risk-taking and behavior within financial systems [20], while cyber insurance can reduce risks associated with digital and international trade [13]. Cyber insurance has also been discussed as an instrument for reducing legal risks in digital commerce [14]. Although these studies focus on different domains, they support the broader view that experience with one form of insurance can shape perceptions of risk protection and the perceived legitimacy of other insurance products.

The contribution of prior insurance ownership is also relevant to customer trust and organizational reputation. Insurance purchasing requires customers to believe that the insurer will honor long-term commitments, manage funds responsibly, and provide reliable service. Corporate governance and risk-management quality may influence this confidence and, consequently, customers' willingness to expand their relationship with an insurer. Research has linked governance and risk-management practices to the performance and stock returns of insurance companies [3]. Corporate social responsibility may similarly provide reputational protection during cybersecurity incidents, reinforcing stakeholder trust under conditions of uncertainty [15]. Customers with positive previous insurance experiences may therefore be more receptive to purchasing an additional life insurance policy.

In contrast, occupation, income, and educational attainment had limited importance in the final model. This does not necessarily imply that these variables are theoretically unrelated to life insurance demand. Their low predictive contribution may instead reflect the composition of the dataset, which included a large number of employees from major organizational clients with relatively similar socioeconomic characteristics. Machine-learning variable importance depends partly on variation within the data; predictors with restricted ranges or highly homogeneous categories provide less information for distinguishing classes. Moreover, recorded income and occupation may not fully represent household wealth, informal earnings, job security, debt obligations, or access to family resources. The low importance of these variables should therefore be interpreted as a characteristic of the present dataset rather than as conclusive evidence that socioeconomic status is irrelevant to repeated life insurance purchasing.

The findings also have implications for actuarial practice and organizational risk management. Although the model focused on customer behavior rather than underwriting losses, its outputs can support actuarial and managerial decision-making by improving estimates of product demand and identifying promising customer segments. Actuaries play a central role in evaluating financial risk, liabilities, pricing, and long-term solvency [17]. Machine-learning predictions can complement actuarial analysis by providing behavioral information about customer acquisition and cross-selling. At the portfolio level, insurers must also optimize investments and control dependencies among financial assets, as illustrated by the use of copula functions and value-at-risk approaches in insurance portfolio optimization [18]. Reinsurance decisions similarly require careful consideration of dependent risks and appropriate stop-loss thresholds [19]. Together, these areas show that sustainable insurance management depends on integrating customer-level analytics with actuarial, investment, and enterprise-level risk controls.

The results should also be considered within the broader sustainability and resilience challenges confronting insurance companies. Sustainable insurance requires financial strength, efficient operations, customer relevance, responsible data use, and the capacity to respond to emerging environmental and technological risks [23]. Organizational resilience is particularly important during novel and severe crises because conventional risk-management assumptions may become inadequate when conditions change rapidly [21]. A predictive model based on real customer data can improve resilience by enabling an insurer to identify stable sources of demand, allocate marketing resources more efficiently, and adapt product strategies to different customer segments. Nevertheless, the use of machine learning must be supported by reliable accounting information systems, internal controls, and high-quality data-management procedures. Information-system quality and internal control mechanisms are closely connected to organizational performance and the reliability of managerial processes [12].

Overall, the study demonstrates that the purchase of a second endowment life insurance policy can be predicted with a meaningful degree of accuracy using administrative data available at the time of the first purchase. The model's strong ROC–AUC, high recall, stable cross-validation performance, and interpretable importance rankings support its potential use in insurance marketing and customer relationship management. The results also indicate that repeated purchasing is shaped more strongly by age, geographical location, gender, marital status, and prior insurance behavior than by the recorded occupational, educational, and income variables in the present dataset. These findings contribute to the emerging literature on intelligent insurance analytics by connecting policyholder risk aversion, actual purchasing behavior, administrative insurance data, and machine-learning prediction.

The present study had several limitations. First, the data were obtained from one insurance company and its associated technology platform, which may restrict the generalizability of the findings to customers of other insurers or countries with different regulatory, economic, and cultural conditions. Second, the retrospective administrative data were collected for operational rather than research purposes, and some variables were incomplete, inconsistent, or limited in detail. Third, risk aversion was inferred from the purchase of a second endowment life insurance policy rather than measured directly through a validated psychological or economic instrument. Repeated purchasing may also reflect marketing exposure, customer satisfaction, agent influence, organizational employment arrangements, or product accessibility. Fourth, potentially relevant predictors such as number of dependents, household wealth, financial literacy, customer satisfaction, premium amount, payment regularity, agent characteristics, and attitudes toward inflation were unavailable. Finally, the moderate precision indicates that a substantial proportion of customers identified as potential purchasers may not complete a second purchase, and the model's practical value therefore depends on the costs and benefits associated with marketing interventions.

Future studies should validate the model using data from multiple insurance companies, different geographical regions, and longer observation periods. Comparative analyses could examine whether XGBoost outperforms logistic regression, random forests, CatBoost, LightGBM, support vector machines, and neural networks under identical data conditions. Subsequent studies should combine administrative records with survey-based measures of risk aversion, financial literacy, trust, satisfaction, and perceived economic uncertainty to distinguish psychological risk aversion from other motivations for repeated purchasing. Researchers should also investigate the timing of the second purchase through survival analysis or time-to-event machine-learning models. The inclusion of policy characteristics, premiums, payment histories, customer-agent interactions, marketing contacts, household composition, and regional economic indicators may improve predictive accuracy. Model explainability techniques such as SHAP analysis could be used to identify both the direction and magnitude of each predictor's effect at the individual and overall levels. External, temporal, and prospective validation should also be conducted before broad operational deployment.

Insurance companies should integrate predictive models into customer relationship management systems to rank policyholders according to their estimated likelihood of purchasing additional life insurance. Because the model demonstrated relatively high recall but moderate precision, predictions should be used to prioritize leads rather than automatically classify customers as purchasers or nonpurchasers. Marketing teams should develop differentiated campaigns based on age, geographical location, gender, marital status, and previous insurance ownership. Regional differences should inform product design, communication channels, premium structures, and the allocation of agents. Customers who already hold automobile or other insurance policies may be suitable candidates for cross-selling and bundled coverage. Insurers should continuously retrain the model as new data become available, monitor changes in performance, audit predictions for unfair bias, and protect customer confidentiality. Predictive outputs should be combined with managerial judgment, actuarial expertise, customer consent, and transparent communication to ensure that machine-learning applications improve both organizational performance and customer value.

Authors' Contributions

Authors equally contributed to this article.

Ethical Considerations

All procedures performed in this study were under the ethical standards.

Acknowledgments

Authors thank all participants who participate in this study.

Conflict of Interest

The authors report no conflict of interest.

Funding/Financial Support

According to the authors, this article has no financial support.

References

[1] E. Shams and M. Rahimpour, "The role of risk management and its evaluation in the performance of insurance companies," in *14th National Conference on Economics, Management and Accounting*, Shirvan, North Khorasan, Iran, 2022.

[2] R. Adeel, "Navigating Uncertainty: Assessing the Impact of Financial Risk Management on Organizational Performance: Evidence From Insurance Companies in Pakistan," *JPR,* vol. 12, no. 4, pp. 76-83, 2023, doi: 10.61506/02.00129.

[3] A. Samani, S. Mohagheghi, E. Kalateh, and R. Hossein Pour, "Investigating the Impact of Corporate Governance in Insurance and Takaful on the Stock Returns of Insurance Companies Listed on the Tehran Stock Exchange Considering the Role of Risk Management," in *Twentieth International Conference on Research in Management, Economics, and Development*, 2023. [Online]. Available: https://civilica.com/doc/2009878.

[4] S.-M. Jiang, F. Xiang, and I. Yang, "Sensation Seeking and Automobile Insurance Coverage Decisions: A Moderated Mediation Model of Gender and Risk Perception," *Social Behavior and Personality an International Journal,* vol. 51, no. 11, pp. 1-9, 2023, doi: 10.2224/sbp.127774.

[5] J. D. Cummins and X. Xie, "Artificial intelligence and cost efficiency in insurance markets," *Journal of Risk and Insurance,* vol. 90, no. 2, pp. 455-482, 2023. [Online]. Available: https://doi.org/10.1111/jori.12389.

[6] S. M. Shah Amirian, "Application of Artificial Intelligence and Big Data Analytics in Risk Management and Improving Accounting Transparency in Insurance Companies: A Case Study of Pasargad Insurance Company," in *Fourth International Congress on Management, Economics, Humanities, and Business Development*, 2025.

[7] I. S. Gómez and O. M. Pineda, "What is an InsurTech? A scientific approach for defining the term," *Risk Management and Insurance Review,* vol. 26, no. 2, pp. 125-173, 2023, doi: 10.1111/rmir.12243.

[8] A. Braun and R. Jia, "InsurTech: Digital technologies in insurance," *The Geneva Papers on Risk and Insurance-Issues and Practice*, pp. 1-7, 2025, doi: 10.1057/s41288-024-00344-x.

[9] L. Pandiri, *The Complete Compendium of Digital Insurance Solutions: Life, Health, Auto, Property, and Specialized Coverage in the Age of AI, Automation, and Intelligent Risk Management*. 2025.

[10] J. J. Jaber, A. S. Alzwi, S. N. M. Alhjahja, and R. A. Omari, "Understanding the Influence of Technology Risks on Insurance Companies Performance Through InsurTech Innovation: An Empirical Study of Jordanian Insurance Companies," 2026.

[11] H. Fukunishi and Y. Kobayashi, "Care-needs level prediction for elderly long-term care using insurance claims data," *Informatics in Medicine Unlocked,* vol. 41, p. 101321, 2023/01/01/ 2023, doi: 10.1016/j.imu.2023.101321.

[12] E. Riupassa, "Impact of Accounting Information Systems, Work Motivation, and Internal Controls on Employee Performance: A Study at PT. Bank Danamon, Ambon Branch," *Journal of Corporate Governance Insurance and Risk Management,* vol. 10, no. 2, pp. 118-124, 2023, doi: 10.56578/jcgirm100202.

[13] Y. Chen and H. Zhao, "Cyber Insurance as a Tool for Mitigating International Trade Risks," *Journal of Risk and Insurance,* vol. 90, no. 1, pp. 55-82, 2023.

[14] Y. Jafari, "Cyber Insurance and the Reduction of Legal Risks in Digital Trade," *Journal of Law and Technology,* vol. 2, no. 1, pp. 67-90, 2023.

[15] V. Bamiatzi, M. Dowling, F. Gogolin, F. Kearney, and S. Vigne, "Are the Good Spared? Corporate Social Responsibility as Insurance against Cyber Security Incidents," *Risk Analysis,* vol. 43, no. 12, pp. 2503-2518, 2023.

[16] Z. Shasha, "Analysis of Strategies to Enhance the Competitiveness of Financial Insurance Companies," *Financial Engineering and Risk Management,* 2023, doi: 10.23977/ferm.2023.060506.

[17] S. A. Jarallah and K. Bougatef, "The Role of The Actuary in Assessing Financial Risks: a Case Study of Iraqi Insurance Companies," *Journal of Law and Sustainable Development,* vol. 11, no. 11, pp. 22-65, 2023, doi: 10.55908/sdgs.v11i11.2265.

[18] A. Ghodrzi, R. Tehrani, and A. Soori, "Optimizing Investment Portfolios of Insurance Companies Using Copula Functions and Value at Risk Approach," *Investment Knowledge,* vol. 13, no. 49, pp. 325-352, 2024. [Online]. Available: http://www.jik-ifea.ir/article_22064.html?lang=en.

[19] A. N. Mandia, "Stop-Loss Reinsurance Threshold for Dependent Risks," *Journal of Mathematical Finance,* vol. 13, no. 03, pp. 304-320, 2023, doi: 10.4236/jmf.2023.133019.

[20] Y. Chang, X. Yu, W. Shan, F. Wang, and Y. Tao, "Impact of Financial Market Development, Financial Crises and Deposit Insurance on Bank Risk," *Prague Economic Papers,* 2023, doi: 10.18267/j.pep.820.

[21] P. Dahmen, "Organizational Resilience as a Key Property of Enterprise Risk Management in Response to Novel and Severe Crisis Events," *Risk Management and Insurance Review,* vol. 26, no. 2, pp. 203-245, 2023, doi: 10.1111/rmir.12245.

[22] S. Zbar, Ibrahim, HK, "The Impact of Digital Insurance in Reducing the Risks of Climate Fluctuations: In a Sample of Iraq Insurance Companies," *Journal of Information Systems Engineering and Management,* vol. 9, no. 1, p. 25040, 2024, doi: 10.55267/iadt.07.14316.

[23] M. Eling, "Is the insurance industry sustainable?," *The Journal of Risk Finance,* vol. 25, no. 4, pp. 684-703, 2024, doi: 10.1108/JRF-12-2023-0314.

[24] A. Chen, "Firm Performance Pay as Insurance Against Promotion Risk," *The Journal of Finance,* vol. 79, no. 5, pp. 3497-3541, 2024, doi: 10.1111/jofi.13379.