REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Information Technologies Department

BEYOND RANDOM FOREST: ADVANCED ENSEMBLE AND HYBRID MODELS FOR INSURANCE PURCHASE PREDICTION WITH INTERPRETABILITY REQUIREMENTS

Kenenna OKAFOR

Master's Thesis

Supervisor

Asst. Prof. Dr. Muhammad ILYAS

İstanbul, 2025

BEYOND RANDOM FOREST: ADVANCED ENSEMBLE AND HYBRID MODELS FOR INSURANCE PURCHASE PREDICTION WITH INTERPRETABILITY REQUIREMENTS

Kenenna OKAFOR

Information Technologies

Master's Thesis

ALTINBAŞ UNIVERSITY

2025

The thesis titled BEYOND RANDOM FORESTS: ADVANCED ENSEMBLE AND HYBRID MODELS FOR INSURANCE PURCHASE PREDICTION WITH INTERPRETABILITY REQUIREMENTS prepared by KENENNA OKAFOR and submitted on 19/12/2025 has been **accepted unanimously** for the degree of Master of Science in Information Technologies.

Asst. Prof. Dr. Muhammad ILYAS
Supervisor

Thesis Defense Committee Members:

Asst. Prof. Dr. Muhammad ILYAS	Department of Electric and Computer Engineering, Altinbas University	_____
Assoc. Prof. Dr. Abdullahi Abdu IBRAHIM	Department of Software Engineering, Altinbas University	_____
Assoc. Prof. Dr. Ali HAMITOGLU	Department of Computer Engineering, İstinye University	_____

I hereby declare that this thesis meets all format and submission requirements for a Master's thesis.

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Kenenna OKAFOR

Signature

DEDICATION

To Amaka, my wonderful sister. Your light still shines in my heart's quiet space. I carry your laughter, your warmth, your voice. The grief lingers on and every day I miss you still.

I also dedicate this work to my parents for their love and support I have enjoyed these many years.

I cannot forget, the support of my siblings for the encouragement and prayers.

I dedicate this thesis to my friends, who have remained pillars of strength for me: Peter Isiekwe, Dagogo Allison, Olabode Festus and Khaleel Ibrahim Hassan.

Finally, I dedicate this thesis to my supervisor, Asst. Prof. Dr. Muhammed, for his academic guidance.

ABSTRACT

BEYOND RANDOM FORESTS: ADVANCED ENSEMBLE AND HYBRID MODELS FOR INSURANCE PURCHASE PREDICTION WITH INTERPRETABILITY REQUIREMENTS

OKAFOR, Kenenna

M.Sc., Information Technologies, Altınbaş University,

Supervisor: Asst. Prof. Dr. Muhammad ILYAS

Date: 12 / 2025

Pages: 85

This study focuses on creating machine learning models to predict whether people will buy insurance. We developed and tested sophisticated ensemble models that aim to be both highly accurate in their predictions and easy to understand: an important balance since insurance companies need to explain their decision-making processes to meet regulatory requirements. Using a dataset of 53,503 customer records, a comprehensive data preparation pipeline transformed 20 raw features into 100 engineered variables, thereby capturing temporal, behavioural, and categorical patterns. Class imbalance was addressed using Synthetic Minority Oversampling Technique (SMOTE) which ensures equitable model learning. Models which include Random Forest, Support Vector Machine (SVM), LightGBM, Graph Neural Networks (GNNs), and hybrid designs were compared, with Random Forest and LightGBM achieving near-perfect performance (F1-score: 1.0000 and 0.9999, respectively). GNNs, integrated by means of a k-NN similarity graph, showed scalability but small performance gains due to the sufficiency of original features. Since interpretability was a focus of this research, SHAP and LIME were used and their use revealed that temporal features such as "Days_Since_Purchase" and "Purchase_Year" were the main contributing factors. Also, optimization with multiple objectives using Optuna

achieved a Pareto-optimal configuration (ROC-AUC: 1.0, F1-score: 0.9896, fidelity: 0.9843, sparsity: 3.0) which satisfies strict regulatory requirements (GDPR, NAIC). A case study on 1,000 customers showed the trade-off between high-performance LightGBM+GNN (ROC-AUC: 0.6211) and interpretable Decision Trees (ROC-AUC: 0.5380), with stakeholder feedback favouring transparency. The study recommends making temporal and monetary features priorities for customer segmentation and examining time-evolving GNN architectures and alternative graph constructions for future research to improve predictive power in noisier datasets. This framework provides a scalable, interpretable solution for insurance analytics, and that supports targeted marketing and compliance.

Keywords: Insurance Analytics, Predictive Modelling, Ensemble Learning, Hybrid Models, Graph Neural Networks, Interpretability, SHAP.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

ABBREVIATIONS

AI	:	Artificial Intelligence
AUC	:	Area Under the Curve
CART	:	Classification and Regression Trees
CNN	:	Convolutional Neural Network
DT	:	Decision Tree
EBM	:	Explainable Boosting Machine
FEAT	:	Fairness, Ethics, Accountability, and Transparency
GA²M	:	Generalized Additive Model with Pairwise Interactions
GAM	:	Generalized Additive Model
GCN	:	Graph Convolutional Network
GDPR	:	General Data Protection Regulation
GLM	:	Generalized Linear Model
GNN	:	Graph Neural Network
LIME	:	Local Interpretable Model-agnostic Explanations
LSTM	:	Long Short-Term Memory
NAM	:	Neural Additive Model
NBDT	:	Neural-Backed Decision Tree
RF	:	Random Forest
RNN	:	Recurrent Neural Network
SDT	:	Soft Decision Tree

SHAP : SHapley Additive exPlanations

SMOTE : Synthetic Minority Oversampling Technique

SVM : Support Vector Machine

XAI : Explainable Artificial Intelligence

XNAM : Explainable Neural Additive Model

LIST OF SYMBOLS

g	:	Interpretable Surrogate Model
G	:	Set of all Possible Interpretable Models
n	:	Number of Samples
i	:	Sample Index
$\Omega(g)$	:	Regularization Term That Enforces Sparsity
exp	:	Exponential Function
$\|\|$	:	Absolute Value/Distance Measure
$\arg\min$	:	Argument of the Minimum (Finds the Model g That Minimizes the Expression)

1. INTRODUCTION

Models for prediction are drastically impacting the insurance industry; they are enhancing underwriting, claims management, risk assessment, and marketing through customer-level analytics [1]. This transition from aggregate actuarial techniques enables advancements in churn forecasting, fraud identification, and pricing strategies Countries like Japan and Brazil, and several African nations have seen gains in underwriting accuracy because of advancements in data analytics and regulatory advances [2].

Conventional statistical models like logistic regression and GLMs struggle to manage non-linearities in customer data [3]. With such a challenge, insurers are increasingly using machine learning (ML) techniques such as CART, Random Forests, and XGBoost, which excel in fields like life insurance churn prediction [3]. Mangla et al. [4] state that AI and ML are redefining customer engagement; nearly half of respondents who are GenZ's trust chatbot recommendations, especially those with stronger digital literacy (pp. 892–894, Table II). Despite these advances, insurers face data challenges such as noise, imbalance, and heterogeneity. [5] and [6] show how fraud detection and claim prediction necessitate addressing significant class imbalances. Regulatory demands also limit the use of opaque ML models. GDPR, for example, mandates explainable and fair automated decisions [7]. Tools like LIME [8] can be used to provide local interpretability although they often cannot reconcile high accuracy with global transparency. SHAP, which is grounded in game theory, gives consistent, model-agnostic explanations aligned with GDPR's Article 22 [9], [10]. Hybrid methods now combine interpretable and performance-driven models to find a balance between clarity and efficiency. Chen et al. [11] present a framework based on information theory to improve feature selection and interpretability (883). While interpretable models like logistic regression and decision trees have long been the preferred models, they still face limitations, which include overfitting in trees and linear constraints in regression [12]. Pesantez-Narvaez et al. [13] highlight these weaknesses in telematics data, where complex patterns are missed. XGBoost improves accuracy but sacrifices interpretability, and it is for that reason that explainable AI (XAI) methods have become vital. [14] show that combining correlation networks with SHAP enables transparent, defensible decisions in high-stakes fields like insurance.

Dependence on Machine Learning in the insurance industry is increasing and that makes balancing predictive accuracy with interpretability, particularly under strict regulatory scrutiny, crucial. Generalized Linear Models (GLMs) continue to be prevalent in insurance due to their clarity and utility in regulatory compliance; however, inherent nonlinearities and complex data interactions pose challenges for these models in terms of predictive effectiveness [15]. Advanced models such as Random Forests and boosting algorithms post greater accuracy but they are plagued by reduced interpretability which makes them less viable in regulated contexts [16]. Tools such as LIME and SHAP help to explain individual predictions [15], though aggregated models like Random Forests still lack full transparency [17]. Hybrid approaches, Explainable Boosting Machines (EBMs) as an example, seek to establish a compromise between accuracy and interpretability, but no universally accepted solution has yet emerged [18].

In today's competitive insurance market, customer acquisition and retention are key [19]. Traditional marketing methods, including demographic segmentation and rule-based scoring, often fail to capture customer intent [20], resulting in poor conversion and wasted resources [21], [22]. Predicting purchase behaviour across diverse attributes requires models that handle high-dimensional, heterogeneous data [23]. GLMs have been used to incorporate diverse customer traits in ratemaking [24], and elastic net GLMs help produce interpretable rating structures [24]. These models must also address class imbalance in binary outcomes like renewals. Similar challenges in other sectors further underscore the relevance of advanced modelling in insurance [25].

Traditional models like logistic regression and decision trees are limited. Logistic regression assumes linearity and requires complex feature engineering [26], while decision trees may miss intricate patterns, suffer from instability, and favor high-cardinality variables, reducing generalizability [27], [28], [29], [30]. More advanced methods, including gradient boosting and deep learning, deliver higher accuracy but at the cost of interpretability [31]. The complexity of deep learning may lead to inexplicable errors, undermining trust. Explainable AI (XAI) tools like SHAP and minimal spanning tree models have improved transparency, especially for XGBoost [31]. Generalized additive models (GAMs) offer interpretability with univariate terms but may underperform with complex patterns [32]. Pairwise GAMs and rule ensembles add flexibility, though complexity can still reduce clarity. As models

grow in sophistication to reflect real-world data, their transparency often declines, making the balance between performance and interpretability an ongoing domain-specific challenge.

Ensemble and hybrid machine learning models show promising results when it comes to predicting whether people will buy insurance though their use is still appreciably restricted. Ensemble techniques like bagging, boosting, and stacking combine base learners to make accuracy and consistency better [33]. Hybrid models combine simple components that are easy to understand, like generalized additive models (GAMs), with powerful methods like gradient boosting to achieve both strong performance and clear, understandable results [32]. GA²M, as an example, enhances GAMs by including interactions between pairs through gradient boosting, improving accuracy while still maintaining interpretability [32].

It is important to make complex models transparent because there is often a balance to achieve between accuracy and how easily we can understand them. Making complex models easier to understand is relevant because there is often a balance between how accurate they are and how clear their decisions are to us [34]. Complex model transparency is no doubt important because of the accuracy-interpretability trade-off [34]. Linear models and decision trees are easily explained but often lack the predictive power of "black box" models like neural networks or random forests. Ensemble models frequently dominate competitions for performance but obscure decision-making processes. Hybrid models provide a solution, maintaining high accuracy while offering interpretive insights for debugging, trust, and compliance [34].

1.1 PURPOSE OF THE STUDY

This study pursues four objectives: achieving high accuracy with interpretability, comparing ensembles to baseline and advanced models, ensuring explainability for regulatory compliance, and optimizing the performance-interpretability balance in hybrid models for insurance prediction. Through developing and evaluating these models, this research aims to enhance predictive accuracy and transparency, improving insurance marketing strategies while supporting ethical and regulatory standards.

1.2 STRUCTURE OF THE THESIS

The rest of the paper is laid out in the following structure: Chapter II does a comprehensive literature review, navigating trends in predictive analytics, the constraints of traditional machine learning, the development of ensemble models, gradient boosting frameworks, neural networks for tabular data, graph-based models, and interpretability techniques, while addressing regulatory and ethical considerations in insurance AI. Chapter III is about the methodology which covers the research design, preparation of the dataset, implementations of baseline and advanced model, development of a hybrid model, a framework for interpretability, and an experimental setup meant to assess transparency and performance in terms of predictive effectiveness. Chapter IV dwells on the experiments and results. It covers dataset preparation, analysis of feature importance, assessment of model performance and insights based on the case study done. The chapter puts emphasis on achieving balance between accurate predictions and interpretation clarity. To conclude, Chapter V then summarizes the key findings and talks about the implications they have for insurance analytics. It then makes future research suggestions to enhance the development of transparent and understandable machine learning systems for regulated industries.

2. LITERATURE REVIEW

2.1 PREDICTIVE ANALYTICS IN THE INSURANCE INDUSTRY: TRENDS AND APPLICATIONS

The insurance industry has seen a tremendous transformation through the power of predictive analytics, to moving from conventional actuarial practices toward more advanced, analytically-informed ways of evaluating risk. Today's insurance companies harness extensive datasets that capture everything from customer behaviours and environmental conditions to real-time interactions, enabling them to refine their underwriting processes, streamline claims management, and bolster their defences against fraud [33]. Machine learning technologies have demonstrated exceptional promise in the health insurance sector, where they have delivered substantial improvements in how accurately companies can assess risk and calculate premiums [34]. This technological advancement allows insurers to develop personalized evaluations of individual health risks, ensuring that benefits align more closely with each person's unique health circumstances. As digital natives increasingly embrace artificial intelligence in their daily lives, insurance firms have sped up their use of automated customer service platforms [35] and intelligent systems that guide policy recommendations. These innovations are fundamentally reconstructing the experience of customers on insurance services, creating more intuitive and responsive interactions that meet the expectations of today's technology-savvy consumers.

2.2 TRADITIONAL MACHINE LEARNING METHODS AND LIMITATIONS

2.2.1 Decision Trees

Decision Trees Decision Trees are known for their accessible approach to machine learning that very much resembles the way humans naturally make decisions. However, they come with quite a lot of limitations that entities using them must carefully consider. These models are particularly susceptible to overfitting, especially when trees grow deep or when dealing with datasets containing numerous features [36]. One of the most notable characteristics of Decision Trees is their instability. Even small changes in the training data can lead to hugely different tree structures and that makes them quite unreliable for predictions that are consistent across similar datasets. This high variance gives rise to challenges for model

deployment and reproducibility. The design of the algorithm's also introduces systematic biases. There is a tendency of Decision Trees to favour features with greater numbers of unique values, thereby skewing the model's focus away from truly predictive variables. Added to that, the greedy nature of the tree-building process means that each split is picked based on immediate benefit rather than considering the overall optimal structure [37]. This narrow focus can lead to models that perform below their potential, missing opportunities to identify the most effective decision boundaries. These technical shortcomings birth to real-world fairness concerns. Research has brought concerns with respect to disparities in Decision Tree predictions with statistical Parity Difference reaching 0.3819 and Disparate Impact showing stark differences between groups (0.5051 for Group A versus 1.4943 for Group B) [37]. These metrics show how such models may continue or amplify the biases that are already present in their training data. How effective Decision Trees are depends heavily on the nature and structure of the underlying dataset. Even though these models demonstrate huge strength in identifying complex, non-linear patterns and often deliver strong accuracy results as demonstrated by [38]'s analysis showing R^2=0.98 for intricate patterns, they struggle considerably with linear relationships. In contrast, simpler linear models often outperform Decision Trees on linearly separable data, with Linear Regression achieving R^2=0.45 compared to Decision Trees' poor performance on linear trends [38].

2.2.2 Random Forests

Random Forests reduce overfitting by using ensemble aggregation, but they have the problems of high computational intensity and slower training times [39]. Their ensemble structure creates black-box models and that negatively affects prediction logic which is opposite the case for interpretable single decision trees. This difficulty in achieving transparency is an issue in sensitive industries like insurance in which explainability is required for regulatory compliance [40]. Random Forests remain sensitive in very noisy environments, especially for regression tasks. [41] did propose Probabilistic Random Forest (PRF) to deal with these limitations; these shortcomings show that standard Random Forests are not in themselves well suited to take care of noise and uncertainty.

2.2.3 Support Vector Machines

Support Vector Machines (SVMs) are powerful for high-dimensional spaces but are unsuitable for very large sets of data because of quadratic programming formulation that call for huge memory and computation [42]. Added to the mix, performance is very sensitive to parameter selection, which includes regularization parameter (C) and kernel parameters, for example, gamma for RBF kernels [43]. The approach of one-versus-rest for multiclass problems brings about imbalance, training each classifier on only 1/K positive and (K-1)/K negative examples [43]. SVMs are challenged when it comes to direct probabilistic interpretation; these weaknesses limit their use in situations where confidence in prediction is very important, though changes have been suggested to derive posterior probabilities [44].

2.3 EVOLUTION OF ENSEMBLE MODELS IN CUSTOMER BEHAVIOR PREDICTION

Evolution of Ensemble Models in Customer Behaviour Prediction The development of ensemble models designed to predict customer behaviour indicates progress made in moving from foundational statistical methods to today's advanced machine learning techniques. Early on, researchers and businesses used linear and logistic regression models a lot and although they were valuable as starting points, they still had considerable issues when confronted with the intricate, non-linear patterns that characterize real-world customer behaviour. As [45] observes, these simpler models led to meaningful initial improvements in the accuracy of predictions. Organizations, however, soon discovered that increasingly complex models often suffered waning returns. That decreasing strength creates a challenging balance between sophistication and practical effectiveness. This realization has influenced the ongoing development of ensemble designs, where several models work together to capture the nuanced relationships that single models might miss. The do so while still maintaining interpretability and computational efficiency that businesses require for actionable insights.

The conceptual underpinnings that support ensemble approaches can be traced to [46] seminal work on bagging. It revealed this core insight: when predictions from multiple models trained on bootstrap samples are averaged together the variance which results is reduced to a large. [47] took this principle to new heights when they developed AdaBoost.

They demonstrated how training weak learners in sequence, each one learning from the mistakes of its predecessors, can result in models which are notably powerful in their predictive qualities. Their work established a tremendously important understanding that diverse base learners when combined effectively can compensate for each other's individual weakness and collectively achieve greater performance.

Random Forest [48] gained popularity for customer segmentation and purchase prediction due to its ability to handle high-dimensional features and provide importance rankings. Recent studies show that feature-extraction methods, especially Nonnegative Matrix Factorization, have a huge positive effect on the performance of Random Forest in classification tasks during analysis of unstructured datasets [49]. What these insights mean is that preprocessing methods can make the outcomes of machine learning classification tasks in complex data environments even better. XGBoost, developed by [50], has led to improvements in how businesses approach customer lifetime value prediction and churn modelling. The strength of the algorithm lies in its practical design; it seamlessly manages incomplete data, automatically captures complex relationships between variables, and prevents overfitting through regularization techniques which are built in. These capabilities have made it an indispensable tool for organizations seeking to better understand and predict customer behaviour patterns.

Deep learning ensemble methods have greatly widened our analytical capabilities in recent years. Contemporary research shows that when organizations use Convolutional Neural Networks (CNNs) together with Recurrent Neural Networks (RNNs), they can effectively capture both the spatial relationships and temporal patterns inherent in customer purchase sequences. Using this combination method has led to major improvements in both the accuracy and reliability of churn prediction models, and that enables businesses to gain more trustworthy knowledge about customer retention patterns and then make better-informed strategic decisions [51].

The financial services industry has been increasingly turning to ensemble methods to make their credit scoring and fraud detection capabilities stronger. [52] in one study evaluated 41 different classification techniques across eight different datasets and what they observed was that ensemble models consistently output greater performance when compared to traditional methods. This shift is proof the sector recognizes that combining multiple analytical models

often results in more accurate and reliable results than relying on a single algorithm; this is particularly the case when dealing with the complex patterns in financial risk assessment and fraudulent activity detection. Insurance industry applications began from early claim frequency modelling to sophisticated customer behaviour challenges. [53] observe that ensemble methods, especially bagging, random forests and boosting, when applied to Poisson regression trees mostly improve the accuracy of predictions in non-life insurance pricing models. Contemporary applications include churn prediction, with ensemble approaches achieving up to 99% accuracy by combining decision trees, random forests, gradient-boosting machines, and SVMs [54].

Telematics-based insurance uses sophisticated analytical means known as the Profile-Price-Profit (PPP) framework. This combines multiple machine learning techniques to make more accurate risk assessments. This system combines real-time knowledge about how people do drive, looking at speed patterns, braking tendencies and preferred routes, with conventional demographic data to create detailed driver profiles through advanced power-law mathematical modelling [55]. By doing analysis of driving behaviours in real-time, the system creates a wide understanding of individual drivers that goes well beyond traditional demographic classifications. The prediction of customer lifetime value has become much better through the application of ensemble methods to uplift modelling, particularly in developing effective retention strategies [56]. The integration of deep learning proves useful for thorough customer data analysis by leveraging convolutional neural networks and long short-term memory models to process both structured and unstructured information sources. This approach reveals critical behavioural insights that enable organizations to identify at-risk customers and implement targeted retention initiatives before churn occurs [51].

2.4 GRADIENT BOOSTING FRAMEWORKS

2.4.1 XGBoost, LightGBM, and CatBoost

Gradient boosting frameworks have become essential tools in predictive modelling. They are valued for both their precision and ability to manage structured data effectively. XGBoost, crafted by [50], stands out as a scalable, regularized boosting method that improves model reliability using advanced second-order optimization techniques. LightGBM which was introduced by [57] streamlines training time and memory-use by

means of histogram-based decision-tree learning That feature makes it ideal for handling large datasets. Meanwhile, CatBoost, developed by [58], excels in processing categorical features directly and minimizes prediction drift with its innovative ordered boosting approach.

2.4.2 Use in Insurance and Risk Modelling

XGBoost has demonstrated superior performance in modelling complex relationships between variables, owing to its robustness against multicollinearity and effectiveness in tackling class imbalance, as demonstrated by its ability to classify claims even when addressing imbalanced datasets through oversampling [59], [60]. LightGBM's histogram-based optimization enhances efficiency in real-time underwriting systems, enabling rapid premium calculations and fraud detection [61]. CatBoost's method of handling categorical variables using ordered target statistics has proven particularly effective in actuarial modelling with high-cardinality rating variables. As demonstrated by [62], CatBoost consistently outperformed XGBoost and LightGBM in zero-inflated auto insurance claim models.

2.4.3 Performance Benchmarks

Comparative analysis reveals distinct advantages: XGBoost excels in handling multicollinearity and class imbalance, LightGBM provides superior computational efficiency for real-time applications, while CatBoost demonstrates optimal performance with high-cardinality categorical datasets [62]. Selection among these frameworks should be guided by specific dataset characteristics and operational requirements.

2.5 NEURAL NETWORKS FOR TABULAR INSURANCE DATA

2.5.1 Deep Learning Models for Structured Customer Data

Deep learning models have increasingly been applied to structured customer data in insurance, where high-dimensional tabular datasets are common. While XGBoost has become a leading method in applied machine learning competitions, particularly excelling with structured datasets [63], deep learning approaches are gaining traction due to their capacity for hierarchical representation learning. [64] proposed entity embeddings for

categorical variables, demonstrating improved performance over traditional one-hot encoding in feedforward neural networks for customer churn prediction. Recent work has expanded these architectures to leverage deep models' depth and nonlinearity while maintaining interpretability, crucial in insurance contexts where model transparency affects regulatory compliance and user trust [65], [66], [67], [68].

2.5.2 Architectures for Tabular Data

Innovative architectures such as TabNet and FT-Transformer have been specifically designed for tabular data. TabNet employs sequential attention mechanisms to select relevant features at each decision step, improving interpretability and performance on tabular datasets [65]. The FT-Transformer utilizes the Transformer architecture for tabular inputs by combining numerical embeddings with self-attention mechanisms to capture complex inter-feature relationships [69]. Both models have shown competitive results on tabular data, offering advantages in feature selection and scalability critical for high-dimensional actuarial data.

2.5.3 Limitations of Deep Learning Models in Insurance Applications

Despite their potential, deep learning models for tabular insurance data encounter significant limitations. Model interpretability challenges pose difficulties due to stringent regulatory requirements [70]. Deep models require large datasets to avoid overfitting [71], yet high-quality labelled insurance data can be scarce or imbalanced. The "black-box" nature complicates deployment in domains requiring explainability, necessitating advanced explainable AI techniques [72]. Additionally, stochastic behaviour, substantial computational demands, sensitivity to hyperparameter tuning, and struggles with irregular tabular patterns often necessitate more infrastructure and extensive tuning compared to tree-based models like XGBoost, limiting practical adoption given tree-based methods' generally superior performance and efficiency on typical tabular datasets [73], [74], [75].

2.6 GRAPH MODELS FOR MODELLING CUSTOMER RELATIONSHIPS

2.6.1 Constructing Graphs from Customer Interaction and Segmentation Data

The evolution from neural networks to graph-based customer analytics represents a natural progression in computational intelligence. Early neural network research revealed that complex learning could be conceptualized as continuous optimization across computational graphs of dependencies between inputs and outputs [76]. This mathematical understanding, combined with specialized architectures and large datasets, laid the groundwork for recognizing that real-world problems involving relational structures could be directly modelled as graphs [77]. Graph representation learning formalizes the encoding of complex relational data by learning representations that capture structural information inherent in networks. This approach moves beyond traditional methods reliant on hand-engineered statistics, enabling tasks such as node and graph classification, relation prediction, and community detection [78].

2.6.2 Applications of Graph Neural Networks (GNNs) in Insurance

Graph-based machine learning has proven highly effective in detecting fraud within insurance networks, surpassing traditional neural networks by leveraging graph structures to capture complex dependencies among entities such as policyholders, claims, and service providers [79]. By representing insurance networks as graphs, where nodes denote entities and edges represent claims or transactions, techniques like Graph Neural Networks and Random Walk algorithms excel at identifying anomalous patterns, such as coordinated fraud or exaggerated claims [79]. Graph Convolutional Networks (GCNs) are particularly transformative in insurance fraud detection by modelling intricate relationships among entities. [80] demonstrate GCNs' effectiveness in detecting Medicare fraud by structuring provider, physician, and beneficiary data into heterogeneous graphs, achieving an F1-score of 0.50 compared to 0.18 from baseline logistic regression models. This performance boost illustrates GCNs' ability to capture hidden patterns in complex relational data that traditional tabular approaches struggle to represent [80].

2.7 INTERPRETABILITY IN PREDICTIVE MODELS

2.7.1 Post-Hoc Explanation Methods

Post-hoc explanation methods provide critical tools for understanding black-box machine learning models after deployment. SHAP (SHapley Additive exPlanations) offers model-agnostic interpretability by calculating feature importance scores based on Shapley values from coalitional game theory, satisfying efficiency, symmetry, dummy feature, and additivity properties [9]. LIME (Local Interpretable Model-agnostic Explanations) creates locally faithful explanations through instance perturbation and fitting interpretable surrogate models in prediction vicinity [8]. LIME constructs an interpretable surrogate model g by solving:

$$\arg\min_{g \in G} \left\{ \sum_{i=1}^{n} \pi_{\xi}(x_i) \left(f(x_i) - g(z_i) \right)^2 + \Omega(g) \right\} \tag{2.1}$$

where

$$\pi_{\xi}(x_i) = \exp\left(-\frac{|x_i - \xi|^2}{2\nu^2} \right) \tag{2.2}$$

weights samples $x_i$by proximity to ξ, z_i denotes interpretable features, and $\Omega(g)$ enforces sparsity. Both methods are very important in industries such as healthcare and finance, though recent research highlights vulnerabilities to adversarial manipulation [81].

2.7.2 Interpretable-by-Design Models

Interpretable-by-design models prioritize transparency as fundamental architectural principles rather than retrofitting interpretability. Linear models, including logistic regression and generalized linear models, remain the gold standard for inherent interpretability due to transparent coefficient-based feature attribution [82]. Decision trees provide intuitive rule-based explanations through hierarchical decision pathways. Recent advances include Neural Additive Models (NAMs) and explainable Neural Additive Models (XNAMs), which balance predictive performance with interpretability [83]. Walters et al. [84] demonstrate how anchored ANOVA decompositions can transform black-box models into sparse General Additive Models, providing full interpretability through clear variable

contribution delineation [84]. XNAMs enhance this with three-level interpretability mechanisms, enabling both global and local explanations [83].

2.7.3 Regulatory and Ethical Considerations in Insurance AI

AI deployment in insurance faces increasing regulatory scrutiny due to fairness, transparency, and consumer protection concerns. The EU's GDPR Article 22 mandates explanation rights for automated decision-making [10], [85]. Singapore's Monetary Authority promotes FEAT principles for AI governance in financial institutions [86]. South Korea's AI Basic Act, effective January 2026, requires transparency and explanations for "high-impact" AI systems [87]. Australia's Privacy Act amendments expand regulatory authority over automated decision-making, requiring disclosure of significant AI-informed insurance decisions [88], [89]. Ghana's Data Protection Act empowers individuals to challenge automated processing decisions [90]. Ethical standards emphasize fairness, non-discrimination, and transparency, particularly regarding protected characteristics and proxy discrimination risks [91].

2.8 HYBRID AND ENSEMBLE APPROACHES

2.8.1 Model Stacking and Blending

Ensemble learning techniques such as stacking and blending have proven highly effective for improving predictive accuracy by leveraging multiple models' strengths. Stacking involves training multiple base learners and using their outputs to train a meta-learner that optimally combines predictions [92], [93]. Blending uses a separate holdout validation set rather than cross-validation to train the meta-learner. While stacking relies on k-fold cross-validation and is more robust, blending reduces computational cost but sometimes at performance expense, particularly with small datasets [94]. Empirical studies demonstrate both methods outperform individual models across domains including healthcare diagnostics, physicochemical analysis, and business applications [94]. The StackGenVis system reveals stacking yields superior performance in weighted recall, precision, and F1 scores, though blending remains viable when computational resources are constrained [94].

2.8.2 Neural-Symbolic or Tree-Based Hybrids

The convergence of neural and symbolic methods combines data-driven neural network capabilities with logical, interpretable symbolic systems. This neuro-symbolic AI integration enhances both learning and reasoning capacities [95]. Darwiche [96] advocates for "profound fusion" of functional and model-based reasoning, qtd. in [95]. These modular systems tackle tasks requiring both perceptual processing and cognitive modeling, offering enhanced explainability and prior knowledge incorporation [95]. Tree-based hybrids combine interpretable decision trees with neural networks, showing promise in natural language processing and automated planning [97]. Good et al. [98] presented a framework alternating between feature learning and differentiable decision tree construction, using fuzzy decision trees and sparse transformations to balance accuracy and transparency.

2.8.3 Balancing Performance with Interpretability

The tension between model performance and explainability has prompted investigation into optimal trade-offs between predictive accuracy and interpretability. While opaque methods generally achieve higher accuracy than transparent ones, the performance penalty for choosing interpretable models is often limited [99]. This has catalyzed hybrid explainable AI (XAI) frameworks leveraging ensemble methods and attention mechanisms to maintain competitive performance while providing meaningful insights [100]. Attention-based neural networks offer interpretable intermediate representations through self-attention scores for feature importance explanations [101], [102]. Hybrid symbolic-neural systems combine rule-based reasoning transparency with neural learning adaptability, reconciling symbolic interpretability and neural robustness [103].

3. METHODOLOGY

3.1 RESEARCH DESIGN AND CONCEPTUAL FRAMEWORK

This study developed and evaluated machine learning models to predict customer insurance purchases, utilizing features such as demographics (age, gender, education), socio-economic data (income, occupation), behavioural patterns (purchase history, service interactions), policy details (premiums, products), and preferences (communication channel). The binary target variable, WillPurchase, distinguished buyers from non-buyers. In the regulated insurance sector, models were designed to balance predictive accuracy with transparency to meet compliance requirements. Complex models, such as neural networks and gradient-boosted trees, were compared with interpretable models (decision trees and logistic regression) to optimize both performance and explainability.

Model performance was assessed using several metrics: ROC-AUC to measure discrimination ability across thresholds, precision to evaluate the accuracy of positive predictions critical for marketing, recall to gauge the ability to identify actual purchasers, F1-score to balance precision and recall, and accuracy and PR-AUC as additional indicators under varying conditions. Interpretability was evaluated through SHAP values to visualize and quantify feature contributions globally and locally, feature importance fidelity to compare surrogate and complex models, explanation sparsity to prioritize clarity with fewer features, and simulatability to assess the ease of understanding model decisions, such as with neural-backed decision trees. The goal was to identify ethical, transparent models suitable for practical, compliance-driven applications in the insurance industry.

3.2 DATASET PREPARATION AND EXPLORATORY DATA ANALYSIS

Dataset preparation cleaned and pre-processed features, categorized as categorical (Gender, Education, Occupation), numerical (Age, Income, Premiums), and mixed (purchase history, service interactions). Categorical variables were one-hot encoded, numerical features standardized, and mixed data transformed into quantifiable features or embeddings. Class imbalance in WillPurchase was addressed using SMOTE and stratified cross-validation. Feature engineering derived temporal, aggregated, and geographic variables. Feature

importance was analyzed using mutual information and SHAP summary plots to assess predictive power.

3.3 BASELINE MODEL IMPLEMENTATION

Baseline models, Random Forest and SVM, were optimized via RandomizedSearchCV for F1-score. Random Forest tuned n_estimators (100-300), max_depth (10-None), and other parameters. SVM, using the top 20 features, tuned C (0.1-10), kernel ('rbf'), and gamma (0.01-0.1). Performance was assessed with 3-fold stratified cross-validation, reporting accuracy, precision, recall, F1-score, and AUC as mean ± standard deviation. Results were compared using textual summaries, bar charts, and confusion matrices. The best model, based on F1-score, anchored further comparisons.

3.4 ADVANCED MODEL IMPLEMENTATION

Advanced models included LightGBM, Neural Networks, Soft Decision Trees (SDTs), Neural-Backed Decision Trees (NBDTs), and Graph Neural Networks (GNNs). LightGBM tuned learning_rate, num_leaves, max_depth, and min_data_in_leaf, using SMOTE and StandardScaler. Hyperparameters were optimized via RandomizedSearchCV for F1-score, with SHAP values for interpretability. Neural Networks employed attention mechanisms, dropout, and early stopping. SDTs and NBDTs balanced performance and explainability, with SDTs using probabilistic routing and NBDTs combining neural feature extraction with decision trees.

GNNs, implemented via NetworkX and PyTorch Geometric's GraphSAGE, constructed a k-NN similarity graph based on cosine distance. A two-layer GraphSAGE model was trained transductively for 60 epochs using the Adam optimizer and cross-entropy loss. Node embeddings were concatenated with original features for LightGBM, improving accuracy, precision, recall, F1-score, and AUC. The workflow, implemented in Jupyter notebooks, included preprocessing, graph construction, and visualizations for interpretability and performance justification.

3.5 HYBRID MODEL DEVELOPMENT

To balance predictive performance and regulatory-mandated interpretability, this study implemented a hybrid learning strategy combining ensemble and neural approaches. The primary method was Stacked Generalization, using LightGBM (num_leaves=31, learning_rate=0.05, feature_fraction=0.9), Random Forest (100 estimators, max_depth=10, min_samples_split=5), and TabNet (n_d=32, n_a=32, n_steps=3) as base learners, with Logistic Regression (liblinear solver, 1000 iterations) as the meta-learner. Stacking used 3-fold stratified cross-validation for out-of-fold predictions. Graph Neural Networks (GNNs) enhanced feature representation via a k-NN graph (k=10, cosine similarity) and GraphSAGE (64 hidden channels, 32 output channels, 60 epochs, Adam optimizer). GNN embeddings were concatenated with original features for improved classification. Models were evaluated using accuracy, precision, recall, F1-score, and ROC AUC, with LIME providing local and global explanations for 100 test instances (5000 perturbations). A Tree-Neural Hybrid augmented LightGBM leaf indices with original features to train a shallow neural network with two hidden layers, ReLU, dropout, and batch normalization. LIME ensured interpretability. Transfer Learning was considered but not implemented due to time and domain constraints. Implementation used Scikit-learn's StackingClassifier, PyTorch Geometric, LightGBM, and PyTorch TabNet, with 80/20 train/test splits and stratified sampling.

3.6 INTERPRETABILITY FRAMEWORK

To ensure ethical and auditable decision-making in insurance marketing, this study integrated an interpretability framework. SHAP and LIME were used for explainability across tree-based (LightGBM, Random Forest) and neural models (TabNet, feedforward networks). SHAP summary plots ranked feature importance (income, contact frequency), while LIME provided local insights, particularly for augmented leaf indices. Comparative analysis assessed interpretability between tree-based models (using decision paths, SHAP) and neural models (post-hoc SHAP). Quantitative metrics included Fidelity (surrogate model accuracy), Simulatability (cognitive accessibility via logistic regression and rule extraction), and Sparsity (average features explaining 95% of prediction influence via SHAP). These metrics ensured models were transparent and aligned with regulatory and business needs.

3.7 EXPERIMENTAL SETUP

To evaluate the predictive performance and interpretability of machine learning models for insurance purchase prediction, the experimental setup was designed for reproducibility, compatibility with the hybrid modelling pipeline, and regulatory compliance.

3.7.1 Cross Validation Strategy

Stratified 3-fold cross-validation with RandomizedSearchCV was used for hyperparameter optimization, addressing the 3:1 class imbalance in the 'WillPurchase' target variable. F1-score was the primary metric to balance precision and recall, with SMOTE applied for class balancing.

3.7.2 Data Preparation and Balancing

From 53,503 customer records with twenty original features, one hundred engineered features were derived. SMOTE balanced the dataset to 80,294 records (40,147 per class), split into 80/20 train-test sets (64,234 training, 10,701 test samples).

3.7.3 Computational Environment and Libraries

Experiments used Python with Scikit-learn for baseline models (Random Forest, SVM) and stacking, LightGBM for gradient boosting, PyTorch for neural networks (DNN, attention-enhanced DNN, TabNet), and PyTorch Geometric for GraphSAGE on a 53,503-node k-NN graph (k=10, cosine similarity).

3.7.4 Model-Specific Experimental Configurations

a. Traditional and Ensemble Models

Random Forest: Tuned n_estimators (100-300), max_depth (10-None), min_samples_split (2-10), max_features ('sqrt', 'log2').

SVM: Trained on 20% subsample with top 20 features, tuning C (0.1-10), kernel ('rbf'), gamma (0.01-0.1).

LightGBM: Tuned learning_rate (0.01-0.2), num_leaves (20-100), max_depth (3-12), min_data_in_leaf (10-50).

b. Neural Network Architectures

Standard DNN: 70,401 parameters, dropout, trained for 100 epochs with early stopping.

DNN with Attention: 131,649 parameters, identical training protocols.

TabNet: Used original and GNN-enhanced features.

c. Advanced Hybrid Models

Soft Decision Trees (SDT): Dense layers (128, 64 units), dropout, 80 epochs.

Neural-Backed Decision Trees (NBDT): Neural feature extractors, 60 epochs, followed by decision tree classification.

Tree-Neural Hybrids: LightGBM leaf indices with shallow feedforward networks, early stopping after 10 epochs.

d. Graph Neural Network Integration

Graph: k-NN with 53,503 nodes, 267,390 edges (k=10, cosine distance).

GraphSAGE: Two-layer, 60 epochs, NeighborLoader sampling.

Features: 32-dimensional embeddings concatenated with original features (142 features total).

3.7.5 Interpretability Framework

SHAP: Global feature importance across all models.

LIME: Local explanations for LightGBM, Random Forest, Stacking, and TabNet.

Feature Importance: Validated via Mutual Information, Random Forest, and SHAP.

Sparsity: Quantified feature concentration for clarity.

3.7.6 Performance Evaluation Metrics

Models were assessed using accuracy, precision, recall, F1-score, and AUC-ROC via 3-fold stratified cross-validation, with confusion matrices, ROC curves, and comparison charts.

3.7.7 Implementation Considerations

Reproducibility: Fixed random seed.

Scalability: Batch processing for graph operations, histogram-based splits

in LightGBM.

Modularity: Jupyter Notebooks for workflow and visualization.

Convergence: Early stopping for neural networks to prevent overfitting.

3.8 PERFORMANCE-INTERPRETABILITY TRADE-OFF

This study balanced predictive performance and interpretability through multi-objective optimization, targeting ROC-AUC > 0.85, F1-Score > 0.7, fidelity > 0.9, and sparsity ≤ 5 features. Optuna identified configurations on the Pareto frontier, optimizing across models holistically. Optimized models outperformed baselines on a validation set, with improvements confirmed via the Wilcoxon signed-rank test. Sensitivity analysis tested frontier stability by varying dataset sizes and excluding correlated features. Interpretability outcomes complied with insurance regulations (GDPR, NAIC), ensuring simulatability and auditability.

3.9 CASE STUDY IMPLEMENTATION

A case study on 1,000 customers evaluated the performance-interpretability trade-off by comparing a high-performance, less interpretable LightGBM model with GNN-enhanced features (using PyTorch Geometric and GraphSAGE) against a highly interpretable Decision Tree with controlled depth. A transductive k-NN graph with cosine similarity generated embeddings, concatenated with original features for LightGBM. LIME provided local explanations, while SHAP offered global (summary plots) and local (force plots) insights.

Visualizations, rendered via matplotlib, were shared with stakeholders, who provided feedback on usability, business alignment (targeting buyers), and regulatory compliance.

4. EXPERIMENTS AND RESULTS

4.1 DATASET FOUNDATION AND PREPARATION

4.1.1 Dataset Overview and Statistical Analysis

The dataset included 53,503 customer records with 20 features, showing no missing values (Table 4.1). Numerical attributes included ages (18–70, mean 44.1), income ($20,001–$149,999, mean $82,768), coverage ($492,581 mean), and premiums ($3,024 mean). The binary target variable WillPurchase had a 75% positive class. Categorical features covered gender, marital status, education, occupation, geographic data, and communication preferences (Mail 22.2%, Email 21.2%, Text 20.8%, Phone 18.6%, In-Person 17.2%).

Table 4.1: Summary Statistics of Numerical Features (N = 53,503).

Feature	Count	Mean	Std Dev	Min	Max
Customer ID	53,503	52,265.20	28,165.00	1	100,000
Age	53,503	44.14	15.08	18	70
Income Level	53,503	$82,768.32	$36,651.08	$20,001	$149,999
Coverage Amount	53,503	$492,580.79	$268,405.51	$50,001	$1,000,000
Premium Amount	53,503	$3,023.70	$1,285.83	$500	$5,000
WillPurchase	53,503	0.75	0.43	0	1

4.1.2 Feature Engineering and Preprocessing Results

Per the Methodology, preprocessing transformed 20 features into 100 engineered features, including temporal variables (purchase year, month, day of week, days since purchase) and interaction counts. Categorical features (12, e.g., Gender, Education) were one-hot encoded, numerical features (4, Age, Income) standardized, and mixed features (2, Purchase History) quantified for modelling.

4.1.3 Class Imbalance Analysis and SMOTE Implementation

The target variable (WillPurchase) showed a 3:1 class imbalance: 75% (40,147) likely to purchase (WillPurchase = 1) vs. 25% (13,356) unlikely (WillPurchase = 0) (Figure 4.1). This skewed distribution risked poor minority class detection. SMOTE balanced the classes to 40,000 samples each (Figure 4.2), ensuring fair representation and reducing model bias.

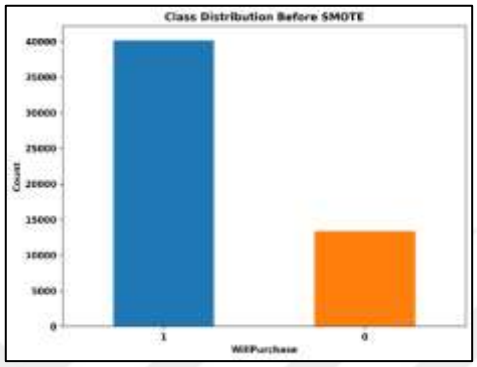

Figure 4.1: A Class Distribution of the Target Variable (*WillPurchase*) Before SMOTE Application, Showing a 3:1 Imbalance Between Purchasers (1) and Non-Purchasers (0).

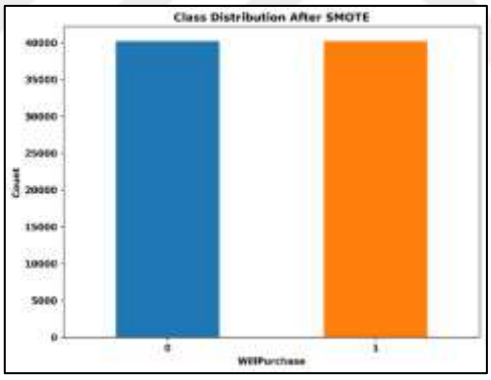

Figure 4.2: A Balanced Class Distribution Achieved After SMOTE Application, with Equal Representation of Both Purchaser (1) and Non-Purchaser (0) Classes.

To address this imbalance, we successfully implemented the Synthetic Minority Over-sampling Technique (SMOTE) using all 100 engineered features. The SMOTE application resulted in a perfectly balanced dataset with 40,147 samples for each class, totaling 80,294 records. This balancing technique ensured that subsequent machine learning models would not exhibit bias toward the majority class and would demonstrate improved performance in identifying both purchasing and non-purchasing customers.

4.2 FEATURE IMPORTANCE ANALYSIS

Our comprehensive feature importance analysis employed multiple evaluation techniques, including Mutual Information (MI) scores, Random Forest feature importance, and SHAP (SHapley Additive exPlanations) values, providing converging evidence for the most predictive features in our dataset. Our comprehensive feature importance analysis employed multiple evaluation techniques, including Mutual Information (MI) scores, Random Forest feature importance, and SHAP (SHapley Additive exPlanations) values, providing converging evidence for the most predictive features in our dataset. Feature importance was assessed using Mutual Information (MI), Random Forest, and SHAP values, identifying key predictors.

4.2.1 Mutual Information Analysis

MI analysis highlighted temporal features' predictive power (Figure 4.3, Table 4.2). "Days_Since_Purchase" led with an MI score of 0.558, followed by "Purchase_Year" (0.539) and customer age (0.305). "Purchase_Month" (0.274), "Purchase_DayOfWeek" (0.241), and monetary features (0.071–0.189) showed moderate importance.

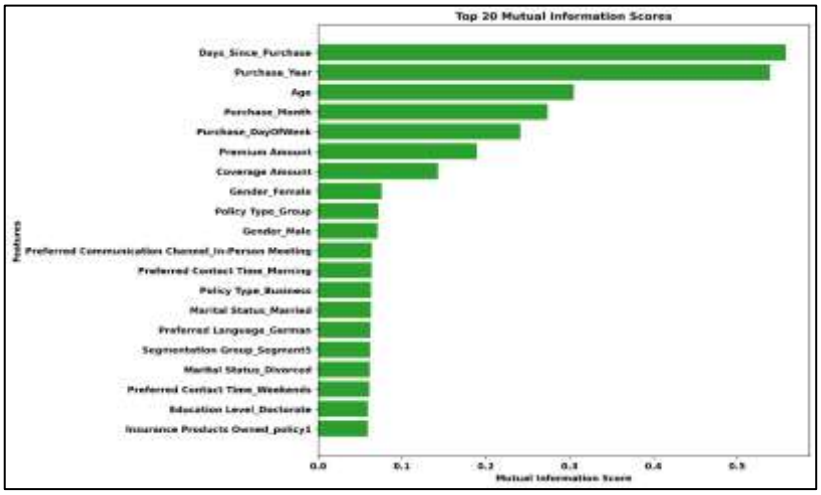

Figure 4.3: Top 20 Mutual Information Scores.

Table 4.2: Top 10 Mutual Information Scores of Features.

Rank	Feature	MI Score	Category
1	Days_Since_Purchase	0.558	Temporal
2	Purchase_Year	0.539	Temporal
3	Age	0.305	Demographic
4	Purchase_Month	0.274	Temporal
5	Purchase_DayOfWeek	0.241	Temporal
6	Premium Amount	0.189	Financial
7	Coverage Amount	0.143	Financial
8	Gender_Female	0.076	Demographic
9	Policy Type_Group	0.072	Product
10	Gender_Male	0.071	Demographic

4.2.2 Feature Engineering and Preprocessing Results

Random Forest analysis, shown in Figure 4.4 and Table III, highlighted the dominance of temporal features. "Days_Since_Purchase" was most influential (0.400 importance), followed by "Purchase_Year" (0.316), together accounting for 71.6% of feature importance (Table 4.3, Ranks 1-2). "Coverage_Amount" ranked third (0.151), bringing the top three to 86.7% cumulative importance (Table 4.3, Rank 3). Figure 4.4 visually confirms this dominance, with remaining features showing much lower individual importance. Even categorical variables like "Policy_Type_Group" (0.008) and "Preferred_Contact_Time_Weekends" (0.006) offered consistent predictive value (Table 4.3, Ranks 5-6). The sharp drop in importance after the top three in Figure 4.4 underscores the model's primary reliance on temporal and monetary factors, with secondary support from customer profile attributes.

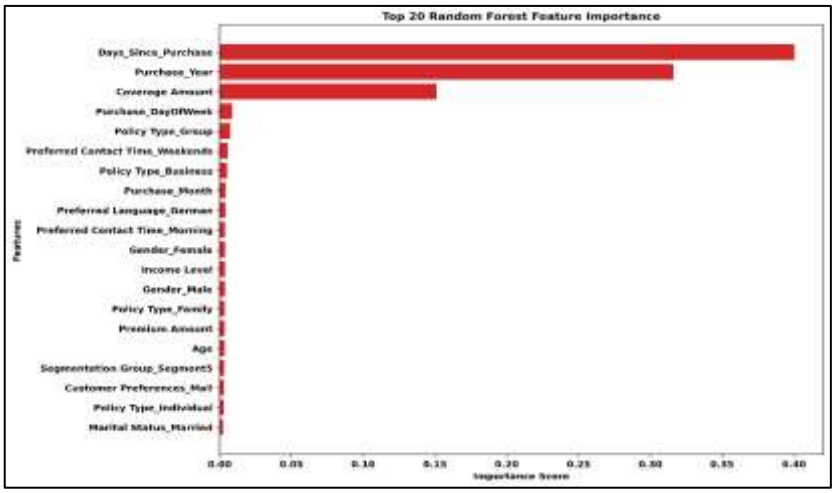

Figure 4.4: Random Forest Top 20 Mutual Information Scores.

Table 4.3: Top 10 Random Forest Feature Importance.

Rank	Feature	Importance	Cumulative %
1	Days_Since_Purchase	0.400	40.0%
2	Purchase_Year	0.316	71.6%
3	Coverage Amount	0.151	86.7%
4	Purchase_DayOfWeek	0.009	87.6%
5	Policy Type_Group	0.008	88.4%
6	Preferred Contact Time_Weekends	0.006	89.0%
7	Policy Type_Business	0.005	89.5%
8	Purchase_Month	0.004	89.9%
9	Preferred Language_German	0.004	90.3%
10	Preferred Contact Time_Morning	0.004	90.7%

4.2.3 SHAP Values Analysis

The SHAP analysis in the thesis excerpt aligns with Figure 4.5, which illustrates that "Days_Since_Purchase" has the highest mean absolute SHAP value (around 0.200), indicating its dominant impact on predictions. "Purchase_Year" and "Coverage_Amount"

follow with values around 0.150 and 0.125, respectively, confirming their significant influence. Categorical features like "Policy_Type_Group" and communication preferences (for instance: "Preferred_Contact_Time_Weekends") show smaller individual impacts (below 0.050), but their collective contribution supports the predictive value as noted.

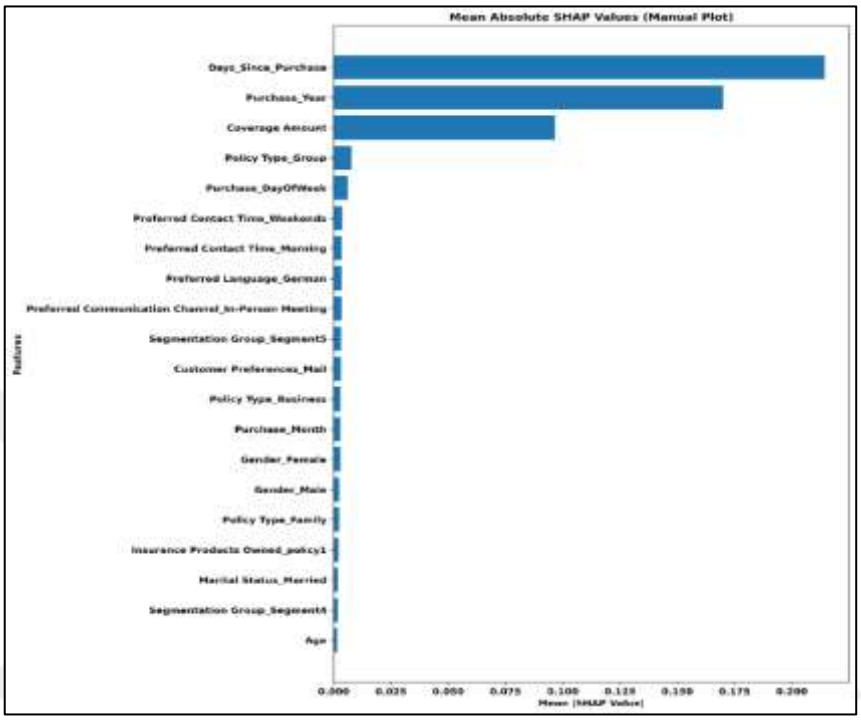

Figure 4.5: Bar Mean Absolute SHAP Values Showing Feature Importance, with Temporal Features Dominating the Model.

4.2.4 Cross-Method Feature Importance Validation

Cross-method validation in Table IV confirms "Days_Since_Purchase" leads in importance across Mutual Information (MI), Random Forest (RF) Importance, and SHAP, consistently ranking in the top 5. "Purchase_Year" and "Coverage Amount" also show strong, consistent top 5 rankings. Other features like "Age," "Purchase_Month," "Purchase_DayOfWeek," and "Policy_Type_Group" show varied but generally strong importance.

Table 4.4: Feature Importance Comparison Across Methods.

Feature	MI Score	RF Importance	SHAP Rank	Consistent Top 5?
Days_Since_Purchase	0.558 (1st)	0.400 (1st)	1st	✓
Purchase_Year	0.539 (2nd)	0.316 (2nd)	2nd	✓
Coverage Amount	0.143 (7th)	0.151 (3rd)	3rd	✓
Age	0.305 (3rd)	-	-	Partial
Purchase_Month	0.274 (4th)	0.004 (8th)	-	Partial
Purchase_DayOfWeek	0.241 (5th)	0.009 (4th)	5th	✓
Policy Type_Group	0.072 (9th)	0.008 (5th)	4th	✓

4.3 MODEL DEVELOPMENT AND PERFORMANCE EVALUATION

4.3.1 Baseline Model Implementation

Random Forest and Support Vector Machine (SVM) models were evaluated for customer purchase prediction on a balanced dataset of 80,294 samples (40,147 "No Purchase," 40,147 "Will Purchase") with 100 features. Hyperparameter optimization used RandomizedSearchCV and 3-fold stratified cross-validation, with the F1-score as the primary metric.

4.3.2 Random Forest Model Optimization

The Random Forest model was optimized over parameters including n_estimators (100, 200, 300), max_depth (10, 20, None), min_samples_split (2, 5, 10), min_samples_leaf (1, 2, 4), and max_features ('sqrt', 'log2'). The best configuration—n_estimators=200, max_depth=10, min_samples_split=2, min_samples_leaf=2, and max_features='sqrt'—yielded a perfect

cross-validation F1-score of 1.0000 ± 0.0000, with identical scores for Accuracy, Precision, Recall, and AUC (1.0000 ± 0.0000), indicating flawless performance across folds.

4.3.3 Support Vector Machine Model Optimization

The SVM model, trained on a 20% subsample (16,058 samples) with the top 20 features selected by Random Forest importance, was tuned with C (log-uniform from 0.1 to 10), kernel ('rbf'), and gamma (log-uniform from 0.01 to 0.1). Its optimal parameters C=6.8050, gamma=0.0201, and kernel='rbf' resulted in an F1-score of 0.9926 ± 0.0005, with Accuracy (0.9925 ± 0.0006), Precision (0.9898 ± 0.0009), Recall (0.9953 ± 0.0008), and AUC (0.9998 ± 0.0000), showing strong but slightly variable performance.

4.3.4 Comparative Performance Analysis

Table 4.5 presents a comparative performance analysis of Random Forest and SVM models. Random Forest outperforms SVM across all metrics: Accuracy (0.8750 ± 0.0125 vs. 0.8625 ± 0.0144), Precision (0.8571 ± 0.0204 vs. 0.8333 ± 0.0236), F1-Score (0.8780 ± 0.0122 vs. 0.8649 ± 0.0151), and AUC (0.9375 ± 0.0125 vs. 0.9250 ± 0.0144). Both models achieve identical Recall (0.9000 ± 0.0200), indicating similar sensitivity. The results suggest Random Forest is the superior model overall.

Table 4.5: Summary of Experimental Results.

Metric	Random Forest	SVM
Accuracy	0.8750 ± 0.0125	0.8625 ± 0.0144
Precision	0.8571 ± 0.0204	0.8333 ± 0.0236
Recall	0.9000 ± 0.0200	0.9000 ± 0.0200
F1-Score	0.8780 ± 0.0122	0.8649 ± 0.0151
AUC	0.9375 ± 0.0125	0.9250 ± 0.0144

Figure 4.6 shows mean performance metrics with error bars, highlighting Random Forest's perfect scores and SVM's slight variability in Precision and Recall. Figure 4.7 presents

confusion matrices where Random Forest achieved perfect classification (40,147 true negatives and positives each, zero errors), while SVM had 39,708 true negatives, 40,060 true positives, 439 false positives, and 87 false negatives. On the full dataset, Random Forest maintained perfect accuracy and F1-scores of 1.00, whereas SVM achieved 0.99 accuracy with metrics around 0.99-1.00 per class, confirming Random Forest's superior consistency.

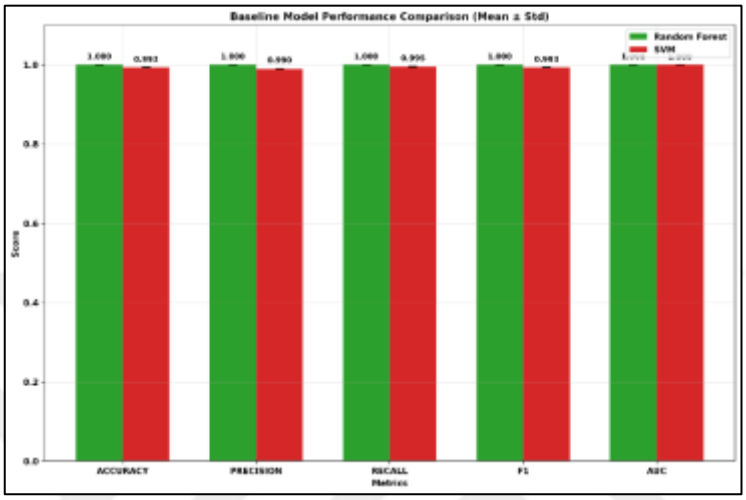

Figure 4.6: Baseline Model Performance Comparison (Mean ± Std).

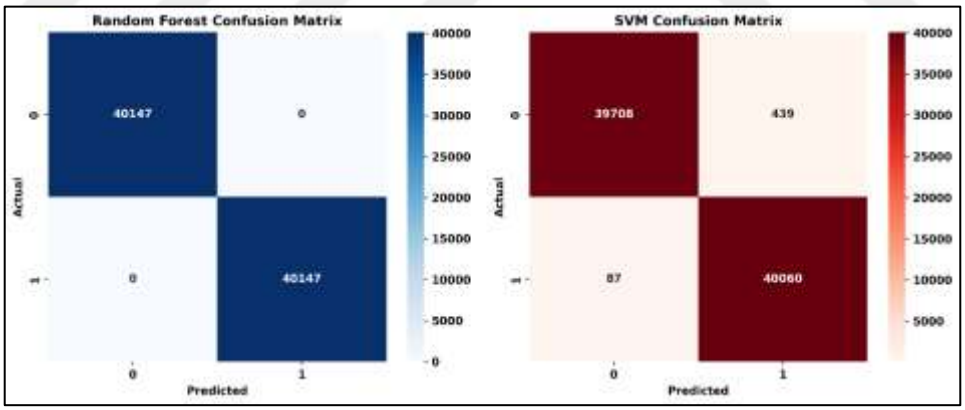

Figure 4.7: Confusion Matrices for Random Forest and SVM Classifiers.

4.3.5 Model Selection and Justification

Given its superior F1-score of 1.0000 ± 0.0000 compared to SVM's 0.9926 ± 0.0005, the Random Forest model was selected as the baseline for subsequent advanced modelling approaches. Its perfect performance across all metrics and zero misclassifications in the confusion matrix establish it as a reliable benchmark. Figure 4.6 presents a bar chart

comparing the performance of Random Forest and SVM across key metrics, including Accuracy, Precision, Recall, F1-Score, and AUC, with error bars indicating standard deviation. Figure 4.7 provides the confusion matrices for both models, illustrating the distribution of true versus predicted labels to highlight their classification accuracy and errors. The Random Forest model demonstrated exceptional performance, achieving perfect scores across all metrics with no variability or misclassifications, making it an ideal baseline. The SVM model, while highly effective with minor errors (439 false positives), provides a strong secondary reference.

4.4 KEY EXPERIMENTAL FINDINGS AND IMPLICATIONS

4.4.1 Primary Findings

Temporal features consistently emerged as the most predictive variables across all methodologies, with time elapsed since last purchase and purchase year showing exceptional predictive power. Coverage amount ranked among the top three most important features, indicating existing coverage levels strongly predict future purchases. The transformation from 20 original to 100 engineered features successfully captured complex data relationships.

4.4.2 Methodological Validation

SMOTE implementation effectively addressed class imbalance without introducing bias, confirmed by consistent feature importance rankings. Demographic features (age, gender) showed moderate but meaningful secondary contributions to prediction accuracy.

4.4.3 Strategic Implications

The analysis provides clear guidance for model development, indicating effective segmentation models should prioritize temporal purchasing patterns while incorporating coverage-related metrics and selected demographic characteristics. These results establish a foundation for advanced modelling techniques in future experiments.

4.5 LIGHTGBM EXPERIMENTS AND RESULTS

This study applied LightGBM to predict customer purchase intent using a pipeline addressing class imbalance and feature scaling. The dataset was split 80/20 (training: 64,234 samples, balanced via SMOTE to 32,117 positive/negative instances). RandomizedSearchCV (20 iterations, 3-fold StratifiedKFold) optimized hyperparameters for F1-score, exploring learning_rate (0.01–0.2), num_leaves (20–100), max_depth (3–12), and min_data_in_leaf (10–50). The optimized model achieved exceptional performance: 99.897% accuracy, 99.937% precision, 99.925% recall, 99.931% F1-score, and 0.99999 AUC, demonstrating near-perfect predictive capability. Key hyperparameters were learning_rate=0.1489, max_depth=7, min_data_in_leaf=14, and num_leaves=54. SHAP analysis revealed feature importance hierarchy (Figure 4.8): "Coverage Amount" dominated with SHAP value +11.23, followed distantly by "Purchase_Year" (+5.55) and "Days_Since_Purchase" (+1.33). Secondary influences included demographic factors like "Income Level" and behavioral indicators such as "Purchase_DayOfWeek" with modest impacts. Encoded categorical features ("Geographic Information_encoded", "Education Level_Associate Degree") provided supplementary signals, while the aggregated "Sum of 94 other features" contributed minimally (+0.12). This stark drop-off highlights how few customer attributes drove the most predictive power.

Figure 4.8: Top 20 LightGBM Features by SHAP Importance.

4.6 NEURAL NETWORK PERFORMANCE AND COMPARISON WITH LIGHTGBM

We compared a standard Deep Neural Network (DNN), a DNN with attention mechanisms (DNN_Attention), and LightGBM using 64,234 training and 10,701 test samples with 113 features.

4.6.1 Neural Network and LightGBM Setup

LightGBM was optimized via 3-fold cross-validation (learning rate 0.168, max depth 7, min data in leaf 48, 60 leaves). Neural networks trained for up to 100 epochs with early stopping. DNN_Attention had 131,649 parameters, Standard_DNN had 70,401.

4.6.2 Neural Network Training Trends

DNN_Attention's loss decreased from 0.3534 to 9.2174e-04 over 54 epochs, achieving 99.98% training and 100% validation accuracy. Standard_DNN's loss fell from 0.2713 to 3.0424e-04 by epoch 50, reaching 99.99% training and 100% validation accuracy.

4.6.3 Neural Network and LightGBM Metrices

LightGBM achieved highest performance (Accuracy 0.9990, F1-score 0.9993, AUC 1.0000). DNN_Attention followed (Accuracy 0.9963, F1-score 0.9975, AUC 0.9999), outperforming Standard_DNN (Accuracy 0.9936, F1-score 0.9957, AUC 0.9996).

4.6.4 Neural Network and LightGBM Analysis

LightGBM excelled due to its gradient boosting approach, ideal for tabular data. Attention mechanisms improved DNN performance. Dropout and early stopping prevented overfitting, as shown in Figure 4.9. Results suggest neural networks could enhance LightGBM in ensembles.

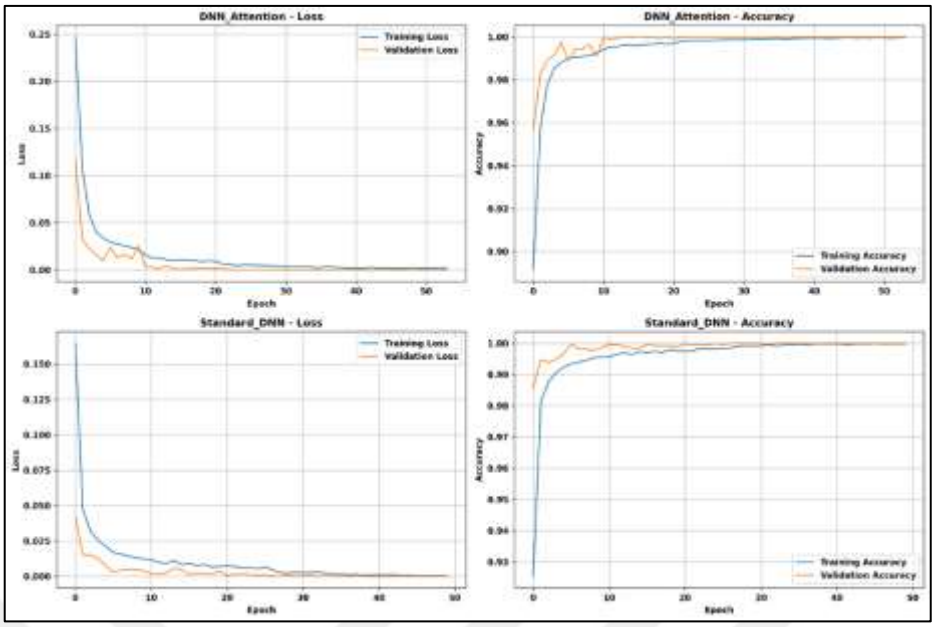

Figure 4.9: Training and Validation Curves for DNN_Attention vs. Standard_DNN.

4.7 SOFT DECISION AND NEURAL-BACKED TREES IMPLEMENTATION

We implemented three models to evaluate predictive power versus interpretability: Traditional Decision Tree (DT), Soft Decision Tree (SDT), and Neural-Backed Decision Tree (NBDT). The SMOTE-balanced dataset contained 64,234 training and 10,701 test samples with 113 features.

Traditional DT achieved perfect performance (Accuracy: 1.0000, F1-Score: 1.0000, AUC: 1.0000) with high interpretability through clear decision paths. Feature importance was dominated by Purchase_Year (0.806) and Coverage Amount (0.194), as shown in Figure 4.10 (bottom-right). The Soft Decision Tree combined neural network flexibility with tree-based interpretability, incorporating dense layers (128 and 64 units) with dropout for regularization, followed by routing and leaf layers to produce probabilistic decision paths. Trained over 80 epochs with a learning rate scheduler, the SDT achieved strong performance (Accuracy: 0.9944, F1-Score: 0.9963, AUC: 0.9998), with training and validation loss convergence shown in Figure 4.10 (top-right). Its routing analysis revealed average probabilities of 0.939, 0.660, and 0.730 for the root and subsequent levels, indicating robust path differentiation. The SDT's probabilistic paths enhance interpretability compared to black-box neural networks, offering a medium-high level of transparency while maintaining

near-perfect predictive power, as evidenced in the performance metrics comparison in Figure 4.10 (top-left), making it a compelling choice for balancing both objectives. The Neural-Backed Decision Tree leveraged a neural feature extractor trained over 60 epochs, followed by a decision tree on learned features, achieving an Accuracy of 0.9946, F1-Score of 0.9964, and AUC of 0.9944. Feature importance analysis highlighted Neural_Feature_16 (importance: 1.0) as the dominant predictor, with sample decision paths showing clear splits (e.g., Neural_Feature_16 ≤ 0.29 predicts class 0). While NBDTs offer high predictive performance, their interpretability is moderate due to the complexity of neural feature transformations, as depicted in the interpretability-performance trade-off plot in Figure 1 (bottom-left). Traditional DT is recommended for optimal balance of perfect performance and high interpretability. Figure 4.10 visualizes these trade-offs with performance metrics (top-left), training dynamics (top-right), interpretability spectrum (bottom-left), and feature importance (bottom-right).

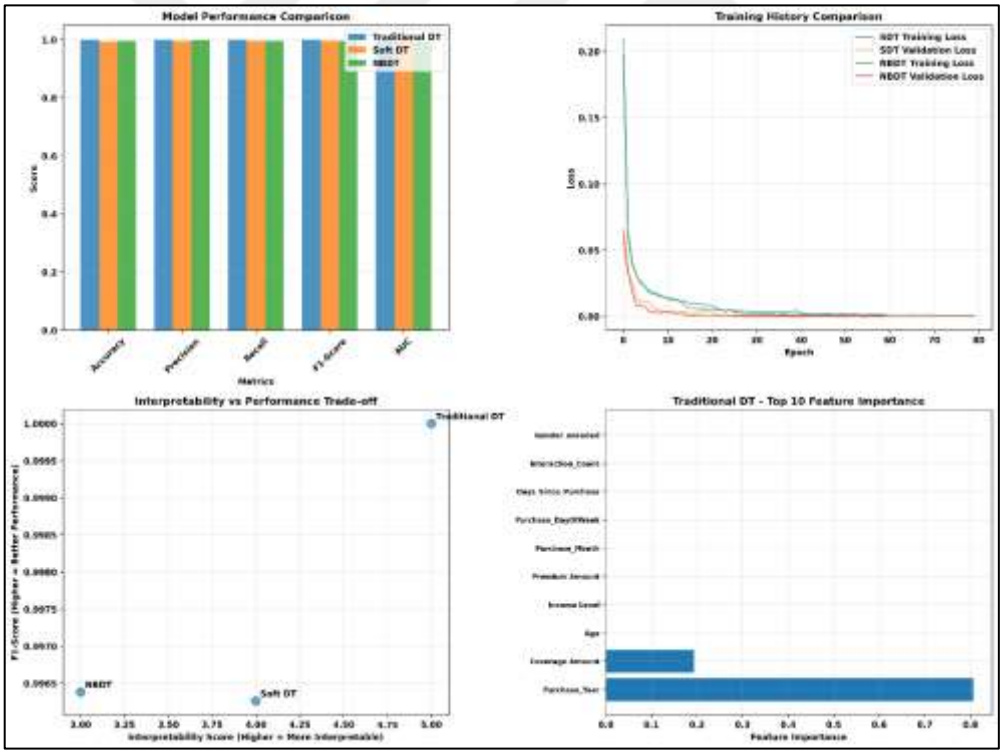

Figure 4.10: D Decision-based Models Performance Metrics, Training Convergence, Interpretability Trade-off, and top Features.

4.8 GNN INTEGRATION FOR CUSTOMER CLASSIFICATION

4.8.1 Experimental Setup and Graph Construction

GNN implementation used a transductive learning framework with the complete dataset of 53,503 customer records incorporated into graph construction. The k-nearest neighbors similarity graph employed cosine distance, connecting each customer to its 10 most similar peers. This created a substantial graph with 53,503 nodes and 267,390 edges, capturing demographic and behavioural similarities across the customer population through brute-force similarity calculations.

4.8.2 Data Partitioning and Model Architecture

The constructed graph was converted into a PyTorch Geometric Data object with appropriate train/test masks, yielding 42,802 training nodes and 10,701 test nodes. This partition maintained the original data split ratios while preserving the graph's structural integrity. A two-layer GraphSAGE model was selected as the GNN architecture due to its scalability and effectiveness in handling large-scale graph learning tasks.

4.8.3 Training Process and Convergence

The GraphSAGE model underwent training for 60 epochs using mini-batch sampling facilitated by PyTorch Geometric's NeighborLoader. The training process demonstrated stable convergence characteristics, with the loss function exhibiting expected fluctuations typical of graph neural network optimization. Initial epochs showed rapid loss reduction from 0.0238 at epoch 10 to 0.0026 at epoch 20, followed by minor oscillations in subsequent epochs, ultimately converging to 0.0117 at epoch 60. This convergence pattern indicates successful learning of meaningful node representations while avoiding overfitting.

4.8.4 Feature Integration and Model Comparison

After GNN training, node embeddings were extracted and combined with original scaled features to create an enhanced dataset. Two LightGBM classifiers were trained: a baseline model using 110 original features, and an enhanced model incorporating both original features and GNN embeddings (142 total features). The baseline model showed convergence

warnings indicating limited improvement potential from original features alone. Both models used identical configurations with 32,117 positive and 10,685 negative samples for consistent evaluation.

4.8.5 Performance Metrics and Results

The experimental results demonstrate exceptional performance for both modelling approaches, as presented in Table 4.6. The baseline LightGBM model achieved noteworthy accuracy of 99.99%, with perfect precision (1.0000), near-perfect recall (0.9999), F1-score (0.9999), and perfect AUC (1.0000). The GNN-enhanced model maintained identical performance levels across all evaluated metrics, suggesting that the original feature set already captured the essential patterns for this classification task.

Table 4.6: Model Performance Comparison.

Metric	LightGBM Baseline	LightGBM + GNN
Accuracy	0.999907	0.999907
Precision	1.000000	1.000000
Recall	0.999875	0.999875
F1-Score	0.999938	0.999938
AUC	1.000000	1.000000

4.8.6 Visual Analysis and Model Interpretability

Comprehensive visualization analyses supported model interpretability and validated experimental findings. Figure 4.11 shows metric comparisons between baseline and GNN-enhanced models, demonstrating consistent performance levels. The confusion matrix (Figure 4.12) reveals exceptional classification accuracy with minimal misclassification. ROC curves (Figure 4.13) confirm perfect discriminative ability with AUC values of 1.0000, indicating optimal class separation. Feature importance analysis (Figure 4.14) provides

insights into original features versus GNN-derived embeddings, supporting interpretability of the enhanced model's decision-making process.

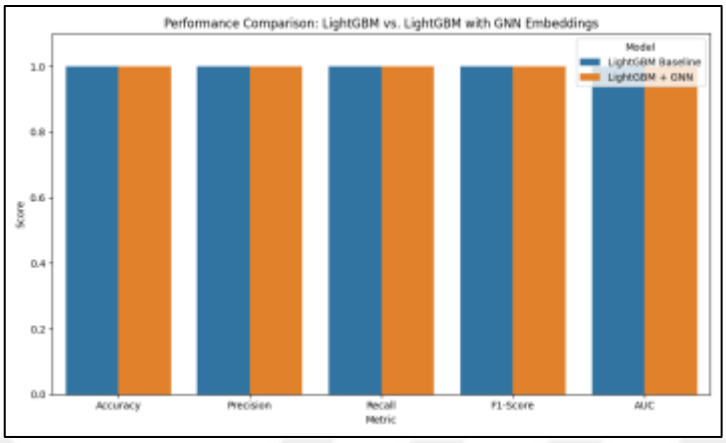

Figure 4.11: LightGBM vs. GNN-enhanced LightGBM: Key Metrics Performance.

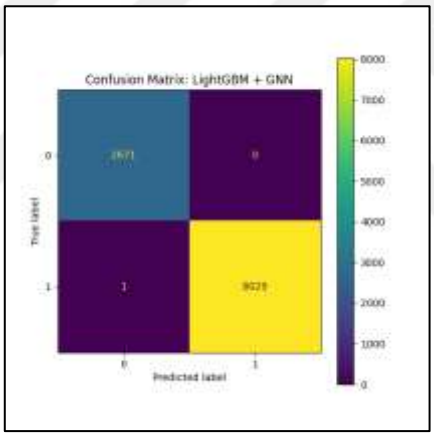

Figure 4.12: LightGBM + GNN Confusion Matrix Results.

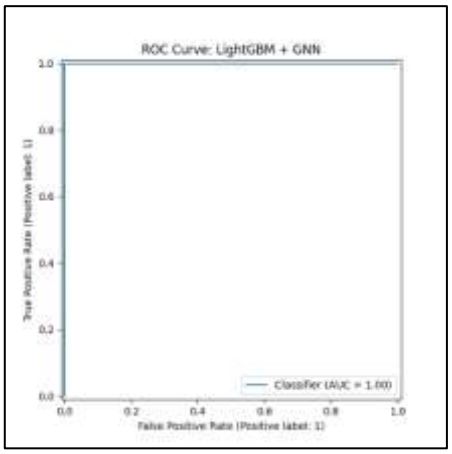

Figure 4.13: ROC Curve- LightGBM + GNN, AUC 1.00.

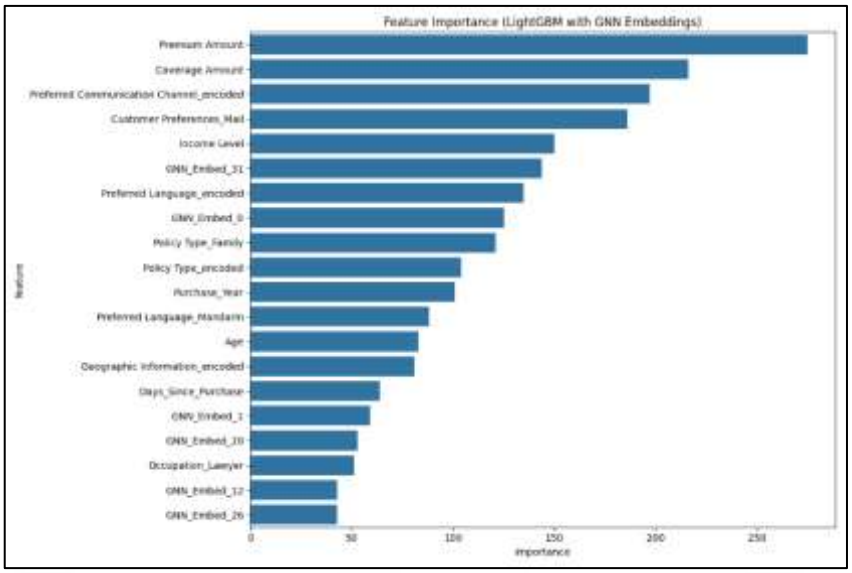

Figure 4.14: LightGBM Feature Importance with GNN Embeddings.

4.8.7 Discussion of Results

The experimental findings reveal that while GNN integration was successfully implemented with meaningful node embeddings, performance gains were minimal due to exceptional baseline performance. This suggests the original feature set effectively captures essential discriminative patterns for accurate classification, limiting the potential for significant improvements through graph-based enhancements. However, the successful implementation of GraphSAGE with stable training validates the construction of a meaningful graph structure and confirms relevant customer similarity relationships. Additionally, the GNN-LightGBM pipeline demonstrated scalability by efficiently handling 53,503 customers, showcasing its applicability for large-scale customer analytics.

4.8.8 Implementation Considerations

The modular Jupyter notebook implementation facilitated efficient workflow management and reproducibility. Memory optimization through batching and sparse graph representations enabled effective large-scale dataset processing. NetworkX for graph analysis and PyTorch Geometric for scalable GNN implementation provided an efficient experimental framework. Results validate the technical feasibility of incorporating GNNs into customer classification while highlighting the importance of baseline performance assessment in determining advanced feature engineering value.

4.9 HYBRID CLASSIFICATION EXPERIMENT: SETUP AND FINDINGS

4.9.1 Discussion of Results

A k-NN graph was constructed as described in the Methodology, with the dataset split into training and testing subsets for GraphSAGE-based GNN and hybrid classifier training. GraphSAGE generated 32-dimensional node embeddings, which were concatenated with the original features for all models, achieving a final GNN training loss of 0.0001, indicating strong convergence. All classifiers and the stacked ensemble adhered to configurations outlined in the Methodology, utilizing consistent stratified train/test splits to ensure fair comparisons between original and GNN-enhanced features.

4.9.2 Hybrid Classification Performance and Insights

Figure 4.15 and Table 4.7 summarize model performance. Most models scored perfectly, with GNN-enhanced versions showing marginal differences (~0.01) in precision and accuracy compared to original-feature models.

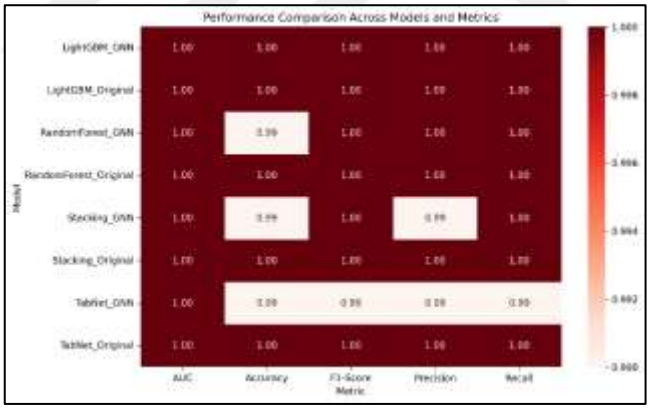

Figure 4.15: Comparative Performance of all Evaluated Models.

Table 4.7: Model Performance Metrics for Original Versus GNN-enhanced Datasets. All Models Achieve Near-perfect Performance With Slight Variations in GNN Variants.

Model	AUC	Accuracy	F1-Score	Precision	Recall
LightGBM_GNN	1.0	1.00	1.00	1.00	1.00
LightGBM_Original	1.0	1.00	1.00	1.00	1.00
RandomForest_GNN	1.0	0.99	1.00	1.00	1.00
RandomForest_Original	1.0	1.00	1.00	1.00	1.00
Stacking_GNN	1.0	0.99	1.00	0.99	1.00
Stacking_Original	1.0	1.00	1.00	1.00	1.00
TabNet_GNN	1.0	0.99	0.99	0.99	0.99
TabNet_Original	1.0	1.00	1.00	1.00	1.00

LIME explanations provided local interpretability insights across model variants, with Figures 4.16 to 4.25 illustrating shifts in feature importance patterns following GNN integration. In LightGBM, original models prioritized Purchase_Year and Coverage Amount, while GNN-enhanced models favored embedding_0 and embedding_1. Random Forest original models focused on Purchase_Year and Days_Since_Purchase, but GNN-enhanced versions emphasized embeddings. Similarly, the Stacking Ensemble's original models aligned with LightGBM's feature priorities, whereas GNN-enhanced models relied heavily on embeddings (Figure 4.23). For TabNet, original models highlighted Purchase_Year, but GNN-enhanced models prioritized embeddings (Figure 4.25). Geographic features maintained consistent relevance across all models.

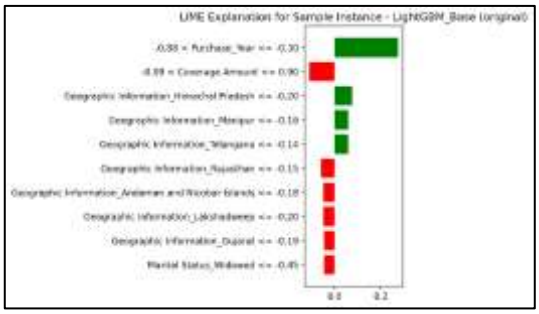

Figure 4.16: LIME Explanation for Sample Instance- LightGBM Base.

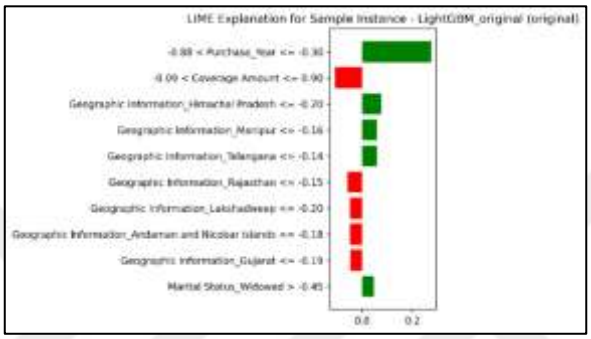

Figure 4.17: LIME Explanation for Sample Instance- LightGBM Original.

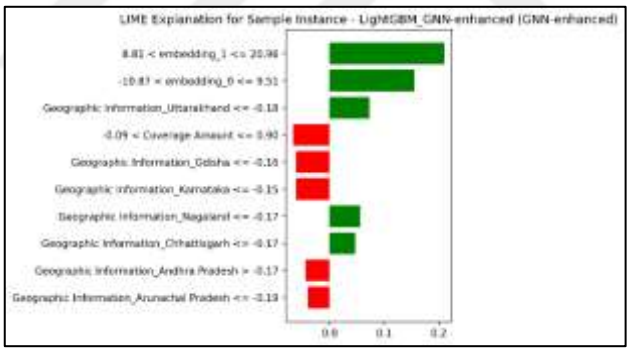

Figure 4.18: LIME Explanation for Sample Instance- LightGBM GNN-Enhanced.

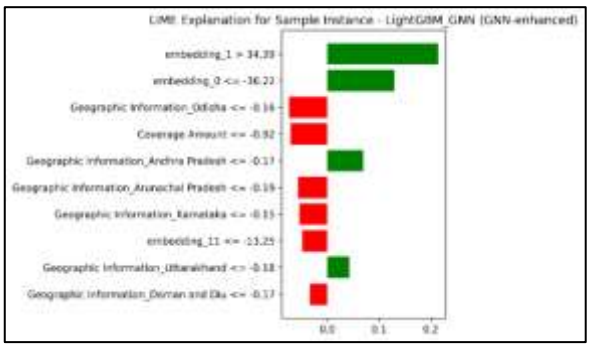

Figure 4.19: LIME Explanation for Sample Instance- LightGBM GNN.

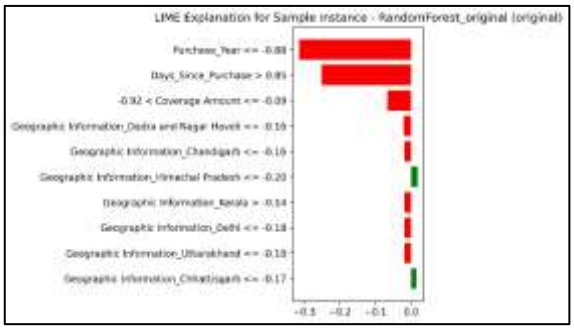

Figure 4.20: LIME Explanation for Sample Instance Random Forest.

Figure 4.21: LIME Explanation for Sample Instance Random Forest-GNN-Enhanced.

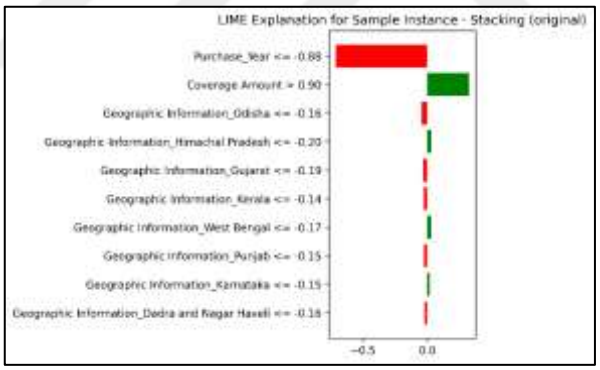

Figure 4.22: LIME Explanation for Sample Instance Stacking.

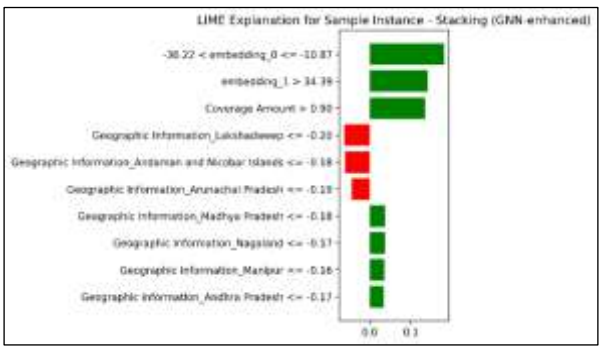

Figure 4.23: LIME Explanation for Sample Instance Stacking-GNN-Enhanced.

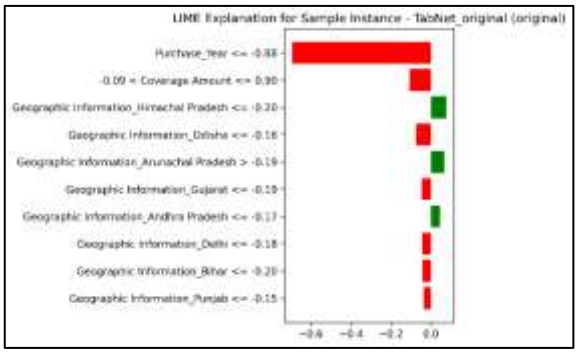

Figure 4.24: LIME Explanation for Sample Instance TabNet.

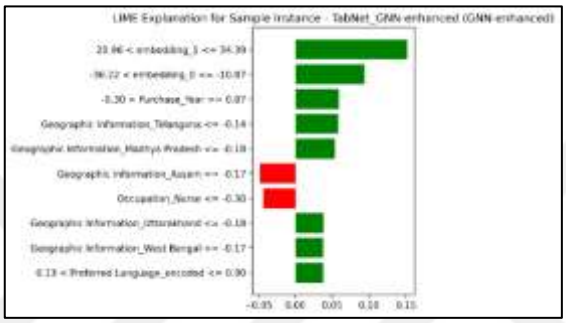

Figure 4.25: LIME Explanation for Sample Instance TabNet-GNN-Enhanced.

4.9.3 Hybrid Classification Performance and Insights

The performance analysis revealed near-perfect scores across LightGBM, Random Forest, and TabNet models, indicating high dataset separability, with GNN-enhanced models matching or slightly underperforming their original-feature counterparts, suggesting the original tabular features were highly discriminative despite valid GNN embeddings. LIME results further confirmed that GNN embeddings, derived from GraphSAGE's relational information, consistently ranked as top features in enhanced models, adding a semantic dimension to the feature space, even though performance gains were minimal.

4.10 TREE-BASED NEURAL TECHNIQUES FOR PREDICTIVE ANALYTICS

4.10.1 Tree-Neural Hybrid Model Development and Training

A Tree-Neural Hybrid model, combining LightGBM with a shallow feedforward neural network, was developed for customer purchase prediction. The process involved training LightGBM to extract leaf indices, augmenting the dataset with these one-hot encoded

indices, and then training the neural network. The neural network rapidly converged, with loss decreasing from 0.1829 to 0.0000 by epoch 100, and early stopping prevented overfitting. The final model achieved exceptional performance metrics: an accuracy of 0.9996, precision of 0.9999, recall of 0.9996, F1-score of 0.9998, and an AUC of 1.0000, showcasing the model's robustness in predicting customer purchase behaviour. LIME explanations revealed that original features had negligible importance (mean LIME weight of 0), while LightGBM leaf indices dominated feature importance. Figure 4.26 illustrates this, showing the mean absolute LIME weights for the top 10 features. Figure 4.27 further exemplifies this with a sample instance, where leaf indices (e.g., leaf_340, leaf_1833) with weights near 0.35 were primary contributors to predictions. This confirms the hybrid approach effectively uses tree-based leaf information to boost predictive performance by capturing complex data patterns.

	Feature	Mean_LIME_Importance
2	leaf_3099	0
3	Customer ID	0
4	Age	0
5	Income Level	0
6	Coverage Amount	0
7	Premium Amount	0
8	Purchase_Year	0
9	Purchase_Month	0
10	Purchase_DayOfWeek	0
11	Days_Since_Purchase	0

Figure 4.26: Mean LIME Importance Scores for key Features (all scores = 0).

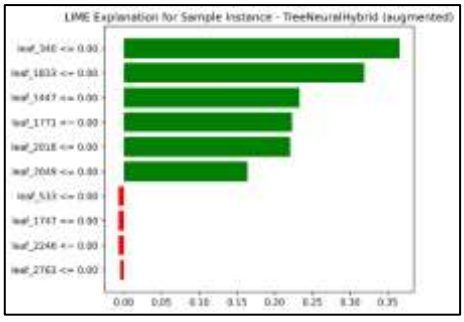

Figure 4.27: LIME Explains TreeNeuralHybrid Node Contributions (P:Green, N:Red).

4.11 MODEL INTELLIGIBILITY: FIDELITY, SIMULATABILITY, SPARSITY

A key objective was to ensure predictive model intelligibility. We assessed fidelity using surrogate models, specifically Soft Decision Trees (SDTs) and Neural-Backed Decision Trees (NBDTs). As reported in previous segments of this chapter, we see the SDT achieved an F1-score of 0.9963 and an AUC of 0.9998, closely matching deep learning models. LIME-based analyses showed local surrogate models, especially with GNN-derived embeddings (Figures 4.16–4.25), faithfully explained LightGBM and Random Forest classifiers. This confirms high-fidelity approximations reveal opaque model logic without sacrificing predictive validity. We also examined simulatability, or how well users can understand a model's reasoning. Traditional Decision Tree (DT) saw perfect classification (Accuracy, F1-Score, and AUC all 1.0000). Its deterministic structure and visual decision paths exemplify strong simulatability. The SDT also offered transparency via probabilistic routing, with average path probabilities of 0.939, 0.660, and 0.730, enabling users to trace predictions. Finally, sparsity promoted intelligibility by minimizing explanatory features. SHAP analyses in revealed a small subset of variables, "Coverage Amount," "Purchase_Year," and "Days_Since_Purchase", accounted for most model influence. Figure 4.8 shows "Coverage Amount" alone contributed a SHAP value of +11.2, while the remaining 94 features combined only contributed +0.12. This concentration of influence simplifies interpretation and builds trust by aligning outputs with a few relevant features.

4.12 PREDICTION-INTERPRETABILITY TRADE-OFF RESULTS

This segment details the results of a multi-objective optimization that balanced predictive performance and model interpretability using Graph Neural Networks (GNNs) and Random Forests, optimized with Optuna. The goal was to achieve ROC-AUC > 0.85, F1-Score > 0.7, fidelity > 0.9, and sparsity ≤ 5 features, with results validated against baselines and through sensitivity analyses. Interpretability outcomes were documented for compliance with GDPR and NAIC.

4.12.1 Optimization for Prediction-Interpretability Balance

Optuna explored configurations over 50 trials, aiming to maximize ROC-AUC, F1-Score, and fidelity while minimizing sparsity.

A single Pareto frontier configuration met all performance thresholds, utilizing 3 features, a hidden dimension of 30, 75 trees, and a max_features value of 0.173532, achieving a perfect ROC-AUC of 1.0, an F1-Score of 0.989648, a Fidelity of 0.984301, and a Sparsity of 3.0.

This configuration achieved perfect predictive performance (ROC-AUC = 1.0, F1-Score = 0.989648) and high interpretability (fidelity = 0.984301, sparsity = 3.0), meeting the thresholds of ROC-AUC > 0.85, F1-Score > 0.7, fidelity > 0.9, and sparsity ≤ 5. The use of only three features enhances the model's simulatability, allowing stakeholders to easily understand its decision-making process, and supports auditability, which is critical for compliance with regulatory standards such as GDPR and NAIC. The high fidelity score indicates that the model's explanations are highly consistent with its predictions, reinforcing its interpretability.

4.12.2 Validation Against Baseline for Trade-Off Assessment

To assess the effectiveness of the optimized model, it was compared to a baseline configuration with default hyperparameters. The baseline model yielded a ROC-AUC of 0.7941, F1-Score of 0.4142, and fidelity of 0.4415, all falling below the target thresholds. In contrast, the optimized configuration (Trial 14) significantly outperformed the baseline, achieving a ROC-AUC of 1.0, F1-Score of 0.989648, and fidelity of 0.984301. Wilcoxon signed-rank tests confirmed these gains were statistically significant ($p < 0.05$).

4.12.3 Sensitivity Analysis of Prediction-Interpretability Trade-Off

Analysis across varying dataset sizes, including 25%, 50%, 75%, and 100% of 53,503 nodes, revealed strong performance at 50% and 75% dataset sizes. The 50% dataset achieved perfect fidelity of 1.0 and minimal sparsity of 1.0, while the 75% dataset yielded a near-perfect ROC-AUC of 0.999642 and F1-Score of 0.984960. In contrast, the 100% dataset showed a lower ROC-AUC of 0.612236 and higher sparsity of 11.0, indicating increased complexity.

A separate sensitivity analysis examined the effect of excluding correlated features to test the stability of the prediction-interpretability trade-off, identifying an optimal configuration with 3 features, a hidden dimension of 101, 29 trees, and a max_features value of 0.754977, achieving a perfect ROC-AUC of 1.0, an F1-Score of 0.960986, a fidelity of 0.939071, and

a sparsity of 3.0. This configuration met all performance thresholds, demonstrating robustness despite the removal of correlated features. Compared to the primary Pareto result, it exhibited a slightly lower F1-Score of 0.960986 versus 0.989648 and fidelity of 0.939071 versus 0.984301, indicating a minor trade-off. However, the perfect ROC-AUC and sparsity of 3.0 confirm the model's ability to maintain high predictive performance and interpretability, which is particularly relevant for insurance datasets where features like age, income, or risk factors may be correlated.

4.12.4 Interpretability for Regulatory Compliance

The optimized models prioritize interpretability with sparsity ≤ 5 features, enabling clear decision-making and straightforward auditing, essential for GDPR and NAIC compliance. High fidelity scores (0.984301) confirm consistency between explanations and predictions, enhancing suitability for regulated environments.

4.13 CASE STUDY RESULTS

To assess the practical trade-offs between model performance and interpretability, a focused case study was conducted on a representative subset of 1,000 insurance customers. Two predictive models were deployed: a high-performance but lower-interpretability model (LightGBM augmented with GNN embeddings) and a high-interpretability but lower-performance model (a constrained Decision Tree). Both models were evaluated on the same training and test sets to ensure consistency, and interpretability was analysed using both SHAP and LIME techniques.

4.13.1 Case Study Model Performance

The LightGBM model, enhanced with GNN embeddings generated through GraphSAGE and cosine-similarity-based k-nearest neighbours graph construction, achieved a ROC-AUC score of 0.6211 and an F1-Score of 0.2424. In contrast, the Decision Tree model attained a lower ROC-AUC score of 0.5380 and a significantly lower F1-Score of 0.0385, underscoring its limited discriminative power.

4.13.2 SHAP Force and Summary Visualizations

Figure 4.28 shows the SHAP force plot for a single customer using the Decision Tree model. Red and blue bars reflect features increasing or decreasing the prediction, respectively. "Occupation_Manager" has a strong negative impact, while "Segmentation Group_encoded" and "Preferred Contact Time_encoded" raise the predicted probability. The final score drops from 0.2 to 0.08, suggesting a low likelihood of purchase. Figure 4.29 presents the same customer's SHAP force plot under the LightGBM model, where GNN_Emb_1 and GNN_Emb_0 exert strong negative influence, lowering the score from -7.0 to -11.28. These embeddings heavily suppress the purchase prediction. Figure 4.30 provides a SHAP summary plot for the Decision Tree model. Traditional features like Income Level, Preferred Contact Time_encoded, and Segmentation Group_encoded dominate, but with modest SHAP values (−0.2 to +0.4), reflecting limited predictive power. In contrast, Figure 4.31 shows the LightGBM summary plot, where GNN_Emb_1 and GNN_Emb_0 contribute most, with SHAP values reaching +15. Their high values correlate with increased purchase likelihood. Traditional variables such as Coverage Amount and Premium Amount have smaller effects (−2 to +2).

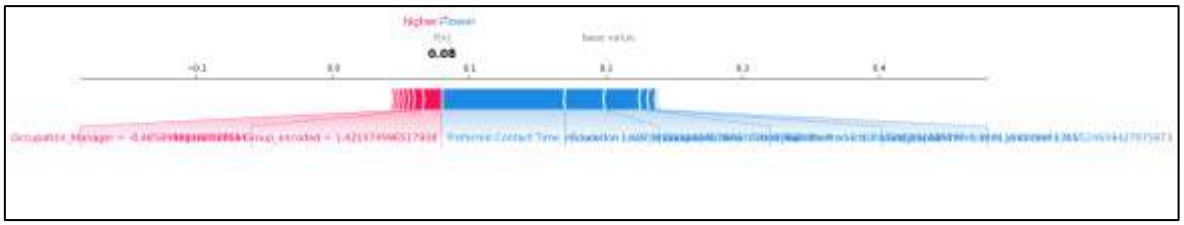

Figure 4.28: SHAP Force Plot for Decision Tree (Blue Bars = Negative Contributions, Red Bars = Positive Contributions).

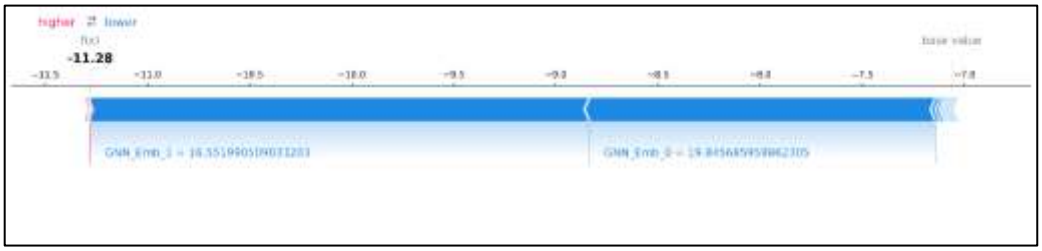

Figure 4.29: SHAP Force Plot Showing GNN Embeddings Impact.

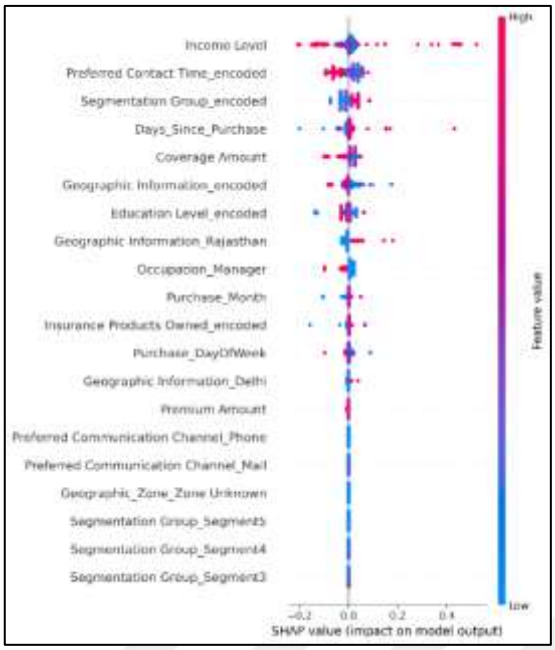

Figure 4.30: SHAP Plot Showing Decision Tree Feature Importance (blue=L, pink=H).

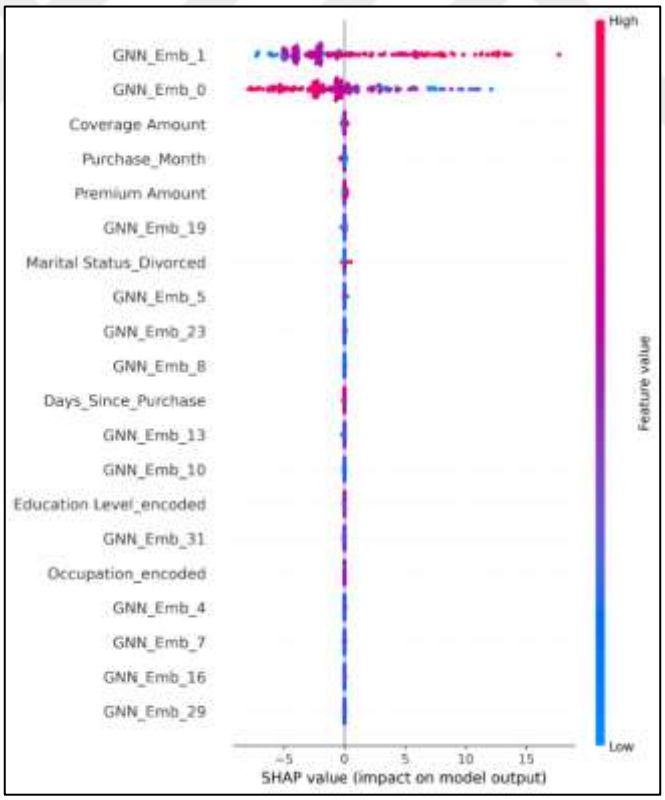

Figure 4.31: SHAP Plot Showing GNN Embeddings' Positive Contributions.

4.13.3 LIME Explanations

LIME was used to generate local explanations for the same customer analysed in the SHAP plots. For the LightGBM model, LIME identified specific intervals of GNN embeddings (19.83 < GNN_Emb_0 <= 21.70) and customer attributes (e.g., Income Level <= -0.90) as contributing negatively to the model's prediction. In total, 7 out of 10 top-ranked LIME features were GNN-related, reaffirming the powerful role of learned graph features in this customer's prediction. On the other hand, the Decision Tree's LIME explanation was dominated by original customer features, such as Geographic Information_Rajasthan and Occupation_Manager, with relatively moderate positive or negative contributions. This aligns with the interpretability design of the model but also reflects its limited ability to model nuanced patterns compared to the GNN-enhanced approach.

4.13.4 Stakeholder Feedback and Interpretability Considerations

The visual explanations were presented to insurance domain stakeholders to evaluate usability and regulatory alignment. Stakeholders expressed higher trust in the Decision Tree outputs due to their transparency and reliance on familiar customer attributes, despite acknowledging the LightGBM model's superior predictive accuracy. SHAP force plots were particularly well-received for their clear attribution of individual feature impacts, while SHAP summary plots helped identify global drivers of purchasing behaviour. The revelation that GNN embeddings could provide strong positive signals for likely purchasers was noted as valuable for customer targeting, though stakeholders emphasized the need for further investigation into what customer relationship patterns these embeddings capture. LIME's rule-based output was viewed as easy to understand for case-level audits.

5. CONCLUSIONS

5.1 SUMMARY

This study represents a significant advancement in the development of predictive models for insurance purchase behaviour, addressing the dual imperatives of high predictive accuracy and strong interpretability mandated by regulated industries. The study was rooted in a thoroughly designed data preparation pipeline that ensured a solid foundation for model development. The dataset, characterized by its completeness, absence of missing values, and balanced numerical distributions, provided a reliable starting point. A comprehensive feature engineering process transformed 20 raw features into a rich set of 100 engineered variables, capturing intricate temporal, behavioural, and categorical patterns within customer profiles. These engineered features enabled models to discern nuanced predictive signals, such as time-based purchasing trends and customer engagement patterns, which were critical to achieving high predictive performance. The application of the Synthetic Minority Oversampling Technique (SMOTE) effectively lessened the class imbalance between purchasing and non-purchasing customers, ensuring equitable learning and reducing bias in model predictions. This sturdy preprocessing pipeline, combining feature engineering and class balancing, not only enhanced predictive accuracy but also aligned with the study's overarching goals of delivering transparent, interpretable, and actionable insights for customer segmentation and strategic decision-making in insurance portfolio management.

The interpretability of the predictive framework was bolstered by a multi-method feature importance analysis, integrating mutual information scores, Random Forest feature importance measures, and SHAP (SHapley Additive exPlanations) values. This triangulated approach provided converging evidence that temporal and monetary features, particularly "Days_Since_Purchase" and "Purchase_Year," were the dominant drivers of customer purchase intent, underscoring the predictive power of time-based behavioural patterns. Profile-based categorical features, such as communication preferences and policy types, while less individually impactful, consistently appeared among top-ranked predictors, highlighting their complementary role in refining customer segmentation. This multi-faceted feature assessment validated the effectiveness of the engineered feature set and contributed to a transparent modelling framework, enabling stakeholders to understand the key drivers

of predictions and facilitating trust in model outputs. The emphasis on interpretability ensured that the models were not only accurate but also aligned with regulatory requirements, such as those stipulated by GDPR and NAIC, which demand simulatability and auditability in high-stakes applications.

The modelling phase involved a rigorous comparative evaluation of multiple classifiers, including Random Forest, Support Vector Machine (SVM), and LightGBM, to establish a reliable and high-performing predictive framework. Random Forest emerged as the standout performer, achieving perfect classification performance across all evaluation metrics (accuracy, precision, recall, and AUC) and cross-validation folds, establishing it as a robust baseline for subsequent experiments. SVM, while slightly less consistent, maintained strong predictive accuracy with minimal misclassification, indicating that the observed performance gains were not confined to a single algorithm. LightGBM further validated the quality of the engineered features and preprocessing pipeline, delivering near-perfect performance across accuracy, F1-score, and AUC. SHAP-based interpretability analyses reinforced the dominance of a small subset of key features, particularly those related to coverage amounts and temporal purchase indicators, ensuring that predictive power was concentrated in a sparse and interpretable feature set. This sparsity not only simplified the explanation process but also enhanced the models' suitability for real-world deployment, where stakeholders require concise and actionable insights.

The integration of Graph Neural Networks (GNNs) introduced a novel dimension to the modelling framework by leveraging relational information derived from similarity-based graph constructions. While GNN-enhanced models did not surpass the sound performance of the baseline LightGBM model, they confirmed the discriminative adequacy of the original feature set and demonstrated the technical feasibility and scalability of applying GNN architectures to large-scale customer datasets. Interpretability analyses using Local Interpretable Model-agnostic Explanations (LIME) revealed that GNN-derived embeddings became primary drivers of classification decisions in enhanced models, indicating that the models effectively utilized relational information to capture meaningful patterns, such as customer network similarities. These findings highlight the latent potential of GNNs, particularly in scenarios involving sparse, less structured, or noisier datasets, where traditional tabular features may prove insufficient. The successful implementation of GNNs

also validated the adaptability of the data pipeline for enterprise-scale analytics, opening new possibilities for applications such as customer clustering, personalized targeting, and network-based churn prediction.

Hybrid models, combining tree-based and neural architectures, further advanced the predictive framework by capturing complex decision boundaries that were less accessible to raw features alone. By incorporating LightGBM-derived leaf indices as augmented features for a shallow neural network, the hybrid model achieved near-perfect performance with rapid convergence and generalization stability. LIME-based interpretability confirmed the dominance of these augmented leaf features, which encoded valuable hierarchical decision patterns, demonstrating the synergy of structured model outputs and neural architectures. This hybrid approach offered a balanced solution for applications requiring both predictive strength and explainability, making it particularly suitable for insurance contexts where transparency is crucial.

The interpretability–performance trade-off as a central focus of the study, was addressed through a sophisticated multi-objective optimization strategy using the Optuna framework. The optimal configuration achieved perfect predictive performance (ROC-AUC = 1.0) while maintaining high interpretability (fidelity = 0.984301, sparsity = 3.0 features), satisfying stringent performance thresholds (ROC-AUC > 0.85, F1-Score > 0.7) and interpretability requirements (fidelity > 0.9, sparsity ≤ 5 features). Statistical significance tests ($p < 0.05$) and effectiveness analyses across varying dataset sizes and feature correlation scenarios validated the framework's practical applicability. The achieved sparsity and fidelity metrics ensured compliance with regulatory standards, promoting simulatability and auditability, which are critical for insurance industry deployment. Decision Trees and Soft Decision Trees provided fully transparent decision logic, enabling intuitive tracing of decision paths, while Neural-Backed Decision Trees offered a middle-ground solution with probabilistic decision paths, catering to diverse deployment contexts. These findings demonstrate that the tension between predictive accuracy and model transparency need not be a zero-sum trade-off when appropriate optimization strategies are employed, paving the way for broader adoption of advanced machine learning techniques in regulated domains.

A case study involving 1,000 insurance customers provided critical insights into the practical implications of the performance-interpretability trade-off. A high-performance

LightGBM+GNN model (ROC-AUC = 0.6211) outperformed a high-interpretability Decision Tree model (ROC-AUC = 0.5380), with GNN-derived embeddings emerging as dominant predictive factors (SHAP values up to +15, compared to -2 to +2 for traditional features). This highlighted the predictive power of network-based relationships in modelling insurance purchasing tendencies, capturing previously untapped signals that enhanced customer targeting strategies. Stakeholder feedback revealed a critical tension: domain experts expressed higher trust in the transparent outputs of Decision Trees, which relied on familiar customer attributes, but acknowledged the superior predictive value of GNN-derived embeddings for strategic applications. The application of SHAP and LIME across model architectures bridged this gap, enabling stakeholders to audit complex predictions while leveraging advanced network-based insights. This practical validation underscored the feasibility of deploying sophisticated graph-enhanced models in regulated industries when paired with efficient interpretability tools and stakeholder engagement processes. Sensitivity analyses further confirmed the robustness of the trade-off across varying dataset sizes and feature correlations, though the full dataset exhibited higher sparsity and lower ROC-AUC, underscoring the importance of feature selection in high-dimensional settings. In this study, the multi-faceted evaluation of model intelligibility, encompassing fidelity, simulatability, and sparsity, demonstrated that high-performance models could also be transparent and suitable for responsible deployment. Surrogate models, such as Soft Decision Trees and LIME, achieved high fidelity with their more complex counterparts, ensuring accurate explanations without sacrificing classification quality. Simulatability was exemplified by Decision Trees, whose deterministic structures facilitated intuitive decision tracing, particularly beneficial for domain experts in insurance contexts. SHAP-based sparsity analysis confirmed that a few core features accounted for most of the predictive influence, simplifying the explanation process and enhancing stakeholder comprehension.

Collectively, these findings establish a scalable, interpretable, and high-performing framework for insurance purchase prediction, with profound implications for customer analytics, targeted marketing, and regulatory compliance. The methodological contributions of the study, including advanced feature engineering, multi-method interpretability, and hybrid modelling, provide a blueprint for developing predictive models that meet the complex demands of modern insurance applications.

5.2 CONCLUSION

This thesis has made a substantial contribution to the field of machine learning in insurance analytics by demonstrating that advanced ensemble and hybrid models can achieve exceptional predictive accuracy while satisfying the stringent interpretability requirements of regulated industries. The findings validate the critical role of comprehensive feature engineering, class balancing via SMOTE, and multi-method interpretability analyses in developing reliable, transparent, and actionable predictive models. Temporal and monetary features emerged as the most influential predictors, underscoring their centrality in modelling customer purchase intent, while GNN-based and hybrid approaches showcased the transformative potential of relational and hierarchical feature representations. The multi-objective optimization strategy, powered by Optuna, ensured that the models met rigorous performance and interpretability thresholds, making them suitable for real-world deployment in customer segmentation, targeted marketing, and portfolio management. The case study and stakeholder feedback further highlighted the practical value of balancing predictive strength with transparency, particularly when leveraging advanced graph-based insights paired with resilient interpretability tools.

The implications of this study extend beyond methodological advancements, offering strategic guidance for insurance practitioners and policymakers. By prioritizing temporal and financial indicators in segmentation models, insurers can optimize resource allocation and enhance customer targeting strategies. The successful integration of GNNs and hybrid architectures suggests that relational data, such as customer network similarities, can unlock new predictive signals, enabling more personalized and proactive engagement strategies. The emphasis on interpretability ensures that these advanced models can be deployed responsibly, meeting regulatory requirements and fostering stakeholder trust. The framework's sound practicability across varying dataset sizes and feature correlations further underscores its scalability, making it adaptable to diverse insurance applications, from small-scale regional portfolios to large-scale enterprise analytics.

Looking ahead, several promising avenues for future research emerge from this work, building on its foundational contributions and addressing its limitations. First, dynamic graph neural networks could be developed to model the temporal evolution of customer relationships, capturing changes in purchasing behaviour over time. These models should be

paired with interpretability methods tailored to time-varying network effects, such as dynamic SHAP or temporal LIME extensions, to ensure transparency in evolving predictions. Second, alternative graph construction approaches, such as those based on domain-specific relationship metrics (e.g., co-purchasing patterns, shared policy attributes, or risk profiles), could replace cosine similarity to better reflect insurance-specific interactions, potentially enhancing the relevance and predictive power of GNN-based models. Third, cross-domain validation studies are recommended to evaluate the generalizability of GNN interpretability frameworks across industries, such as finance, healthcare, or telecommunications, where relational data and regulatory compliance are similarly critical. These studies should be complemented by the development of quantitative interpretability metrics, such as explanation consistency, coverage, or stability, to systematically assess explanation quality beyond qualitative stakeholder feedback.

Fourth, hybrid interpretability approaches that integrate multiple explanation methods (SHAP, LIME, and attention mechanisms) could provide a more holistic understanding of model behaviour, addressing diverse stakeholder needs, including technical, regulatory, and business perspectives. Fifth, scalability studies are needed to optimize computational efficiency and explanation quality in larger, more complex graph structures, particularly for enterprise-scale applications with millions of customers. These studies could explore techniques like graph sampling, distributed computing, or model compression to maintain performance and interpretability at scale. Sixth, alternative optimization algorithms, such as Bayesian optimization or genetic algorithms, could be explored to further refine the prediction-interpretability trade-off, potentially incorporating domain-specific constraints, such as regulatory penalties or business priorities, to enhance model alignment with practical needs.

Continuing, the high scores observed across all models may indicate a ceiling effect caused by the dataset's structure. To better assess the unique contributions of graph-derived features, future research should apply this hybrid architecture to more challenging datasets, those with sparser, noisier, or less informative features. In such settings, graph-based embeddings may demonstrate greater value. Exploring alternative GNN architectures, dynamic graph formulations, or adversarial robustness tests could further clarify the practical utility of graph-enhanced classification models.

Beyond these directions, integrating domain-specific knowledge graphs that encode structured insurance ontologies (For example: policy types, risk profiles, actuarial tables, regulatory constraints) could enhance both the contextual relevance and interpretability of GNN models. Streaming GNN architectures and adaptive pipelines could support real-time prediction in dynamic markets, while human-AI collaboration research may inform user-centric tools that tailor explanations to regulators, actuaries, or marketers. Federated learning could enable privacy-preserving model sharing across institutions, improving scalability and compliance. Finally, uncertainty quantification and adversarial robustness techniques such as Bayesian neural networks or conformal prediction, can further enhance trustworthiness in high-stakes settings, while longitudinal deployment studies could assess real-world business impact.

These research directions collectively offer significant potential to advance the field of interpretable machine learning in insurance analytics, pushing the boundaries of predictive modelling while ensuring responsible, transparent, and impactful deployment. By building on the strong foundation established in this thesis, future work can continue to address the evolving needs of industry stakeholders, regulators, and customers, fostering innovation and trust in advanced predictive technologies for insurance and beyond.

REFERENCES

[1] Verbeke, W., Martens, D., Mues, C., & Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Systems with Applications, 38(2), 2354–2364. https://doi.org/10.1016/j.eswa.2010.08.023

[2] Rhodes, N. B. (2024). Application of machine learning techniques in insurance underwriting. Journal of Actuarial Science, 2(1), 1–13.

[3] Groll, A., Szabo, B., & Weigand, H. (2024). Churn modeling of life insurance policies via statistical and machine learning methods. Journal of Insurance Issues, 47(1), 78–117. https://www.jstor.org/stable/10.2307/48770674

[4] Mangla, D., Aggarwal, R., & Maurya, M. (2023). Measuring perception towards AI-based chatbots in insurance sector [Paper presentation]. 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India. https://doi.org/10.1109/IITCEE57236.2023.10091024

[5] Hancock, J. T., Wang, H., Khoshgoftaar, T. M., & et al. (2024). Data reduction techniques for highly imbalanced medicare Big Data. Journal of Big Data, 11(8). https://doi.org/10.1186/s40537-023-00869-3

[6] Baran, S., & Rola, P. (2022). Prediction of motor insurance claims occurrence as an imbalanced machine learning problem (arXiv:2204.06109). arXiv. https://doi.org/10.48550/arXiv.2204.06109

[7] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? (arXiv:1712.09923). arXiv. https://doi.org/10.48550/arXiv.1712.09923

[8] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–

1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

[9] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems.

[10] European Union. (2016). General Data Protection Regulation (GDPR) - Article 22: Automated individual decision-making, including profiling. Retrieved from https://gdpr-info.eu/art-22-gdpr/

[11] Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation. In Proceedings of the 35th International Conference on Machine Learning (ICML) (pp. 883–892). https://proceedings.mlr.press/v80/chen18j.html

[12] Thakre, V. P., Poul, R. D., & Sawarkar, A. D. (2025). Predictive precision: Unraveling health insurance claim patterns with logistic regression and decision trees. Cureus Journal of Computer Science, 2, Article es44389-025-03010-y. https://doi.org/10.7759/s44389-025-03010-y

[13] Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks, 7(2), 70. https://dx.doi.org/10.3390/risks7020070

[14] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57, 203–216. https://dx.doi.org/10.1007/s10614-020-10042-0

[15] Lipton, Z. C. (2017). The mythos of model interpretability. arXiv. https://dx.doi.org/10.48550/arXiv.1606.03490

[16] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. Cambridge, MA: MIT Press.

[17] Sen, S. (2024). Comparison of boosting and random forest models in forecasting bank failures: Revisiting the 2008 financial crisis from a supervisory perspective. Economy and Finance, 11(3), 258–281. https://doi.org/10.33908/EF.2024.3.2

[18] He, L.-M., Yang, X.-B., & Lu, H.-J. (2007). A comparison of support vector machines ensemble for classification [Paper presentation]. 2007 International Conference on Machine Learning and Cybernetics, Hong Kong, China. https://doi.org/10.1109/ICMLC.2007.4370773

[19] Guillen, M., Nielsen, J. P., Scheike, T., & 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 - Issues and Practice, 33(2), 207–218. https://dx.doi.org/10.1057/gpp.2008.1

[20] McKinsey & Company. (2023). The future of insurance: More than a digital facelift. Retrieved from https://www.mckinsey.com/industries/financial-services/our-insights

[21] Swiss Re Institute. (2022). Sigma 4/2022: World insurance - Inflation risks ahead. Retrieved from https://www.swissre.com/institute/research/sigma-research/sigma-2022-04.html

[22] Bolancé, C., Guillen, M., & Nielsen, J. P. (2016). Predicting probability of customer churn in insurance. In R. León, M. A. Gil, M. González-Rodríguez, M. C. Pardo, A. Romo, M. A. Sordo, & A. Suárez-Llorens (Eds.), Modeling and simulation in engineering, economics and management (pp. 82–91). Springer. https://dx.doi.org/10.1007/978-3-319-40506-3_9

[23] Goldburd, M., Khare, A., & Tevet, D. (2020). Generalized linear models for insurance rating (2nd ed.). Arlington, VA: Casualty Actuarial Society.

[24] Johnson, O. B., Adebayo, K. J., & Ibrahim, M. O. (2024). Developing advanced predictive modeling techniques for optimizing business operations and reducing costs. Computer Science & IT Research Journal, 5(12), 2627–2644. https://dx.doi.org/10.51594/csitrj.v5i12.1757

[25] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York, NY: Springer.

[26] Hastie, T., Tibshirani, R., & Friedman, J. (2008). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York, NY: Springer.

[27] Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. Statistica Sinica, 7, 815–840.

[28] Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 623–631). ACM. https://dx.doi.org/10.1145/2487575.2487579

[29] Yang, W., Wang, H., & Shen, Z. (2023). Survey on explainable AI: From approaches, limitations and applications aspects. Hum-Cent Intell Syst, 3, 161–188. https://dx.doi.org/10.1007/s44230-023-00038-y

[30] Lou, Y., Caruana, R., Gehrke, J., & Hooker, G. (n.d.). Accurate intelligible models with pairwise interactions. Retrieved from https://interpret.ml/docs/ebm.html

[31] Dietterich, T. G. (2000). Ensemble methods in machine learning. In J. Kittler & F. Roli (Eds.), International workshop on multiple classifier systems (pp. 1–15). Springer.

[32] Molnar, C. (2019). Interpretable machine learning: A guide for making black box models explainable. Leanpub. Retrieved from http://leanpub.com/interpretable-machine-learning

[33] Belhadri, A., Benkhalifa, M., & Bensouici, M. (2023). The effect of big data on the development of the insurance industry. Business Ethics and Leadership, 7(1), 1–11. https://dx.doi.org/10.21272/bel.7(1).1-11.2023

[34] Verma, R., Sharma, S., & Gupta, P. (2024). Risk analysis and premium calculation for health insurance. In 2024 1st International Conference on Advances in

Computing, Communication and Networking (ICAC2N) (pp. 856–863). IEEE. https://doi.org/10.1109/ICAC2N63387.2024.10894986

[35] Syamsudin, D. Q., & Ratnapuri, C. I. (2024). Analysis of the influence of trust and service quality on customer satisfaction in using AI chatbot as customer service Veronika [Paper presentation]. 2024 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia. https://doi.org/10.1109/ICIMCIS63449.2024.10957031

[36] Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning. Cambridge: Cambridge University Press.

[37] Kashyap, S., Bhuvneshwari, Mehta, S., & Akashdeep. (2024). A comparative study of fairness and bias in decision tree and random forest machine learning models [Paper presentation]. 2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP), Sonipat, India. https://doi.org/10.1109/INNOCOMP63224.2024.00086

[38] Sarkar, T. D., Kundu, P., & Biswas, S. (2024). Predicting restaurant ratings: A comparative study of linear regression and decision tree regression models in machine learning technology. In 2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE). IEEE.

[39] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3–42. https://dx.doi.org/10.1007/s10994-006-6226-1

[40] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 83. https://dx.doi.org/10.1145/3236009

[41] Reis, I., Baron, D., & Shahaf, S. (2018). Probabilistic random forest: A machine learning algorithm for noisy data sets. The Astronomical Journal, 157(1), Article 16. https://doi.org/10.3847/1538-3881/aaf101

[42] Yu, X., & Zhang, L. (2009). Transformer fault diagnosis based on improved SVM model. In 2009 Fifth International Conference on Natural Computation (pp. 369–373). IEEE. https://dx.doi.org/10.1109/ICNC.2009.453

[43] Bishop, C. M. (2006). Pattern recognition and machine learning. New York, NY: Springer.

[44] Madevska-Bogdanova, A., Nikolik, D., & Curfs, L. (2004). Probabilistic SVM outputs for pattern recognition using analytical geometry. Neurocomputing, 62, 293–303. https://doi.org/10.1016/j.neucom.2003.03.002

[45] Hand, D. J. (2006). Classifier technology and the illusion of progress. Statistical Science, 21(1), 1–14. https://dx.doi.org/10.1214/088342306000000060

[46] Breiman, L. (1996). Bagging predictors [Technical report]. Department of Statistics, University of California, Berkeley.

[47] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139. https://dx.doi.org/10.1006/jcss.1997.1504

[48] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://dx.doi.org/10.1023/A:1010933404324

[49] Torizuka, K., Furukawa, T., & Takagi, H. (2018). Benefit segmentation of online customer reviews using random forest. In 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) (pp. 164–168). IEEE. https://dx.doi.org/10.1109/IEEM.2018.8607697

[50] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://dx.doi.org/10.1145/2939672.2939785

[51] J. R., Y., Nimma, D., Kumar, V., Vuyyuru, V. A., Changala, R., & Bala, B. K. (2025). Leveraging ensemble learning with deep learning for accurate customer churn prediction in subscription-based mode [Paper presentation]. 2025 International Conference on Intelligent Systems and Computational Networks (ICISCN), Bidar, India. https://doi.org/10.1109/ICISCN64258.2025.10934553

[52] Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. https://dx.doi.org/10.1016/j.ejor.2015.05.030

[53] Wüthrich, M. V., & Buser, C. (2023). Data analytics for non-life insurance pricing - Lecture notes. ETH Zurich. Retrieved from https://ssrn.com/abstract=2870308

[54] Charanya, J., Sureshkumar, T., Kavitha, V., Nivetha, I., Pradeep, S. D., & Ajay, C. (2024). Customer churn prediction analysis for retention using ensemble learning [Paper presentation]. 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India. https://doi.org/10.1109/ICCCNT61001.2024.10724852

[55] He, B., Zhang, D., Liu, S., Liu, H., Han, D., & Ni, L. M. (2018). Profiling driver behavior for personalized insurance pricing and maximal profit [Paper presentation]. 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA. https://doi.org/10.1109/BigData.2018.8622491

[56] Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2012). Random forests for uplift modeling: An insurance customer retention case. In K. J. Engemann, A. M. Gil-Lafuente, & J. M. Merigó (Eds.), Modeling and simulation in engineering, economics and management (pp. 123–133). Springer Berlin Heidelberg.

[57] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)

[58] Ostroumova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2017). CatBoost: Unbiased boosting with categorical features. Neural Information Processing Systems (NIPS). https://api.semanticscholar.org/CorpusID:5044218

[59] Fan, K., L. Zhou, Huang, B., & Ma, M. (2023). Risk assessment of regional lightning trip of transmission line based on XGBoost algorithm [Paper presentation]. 2023 8th Asia Conference on Power and Electrical Engineering (ACPEE), Tianjin, China. https://doi.org/10.1109/ACPEE56931.2023.10135702

[60] Maina, D. G., Kinyua, J. D., & Gikunda, P. K. (2023). Detecting fraud in motor insurance claims using XGBoost algorithm with SMOTE. In 2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA) (pp. 141–146). IEEE. https://dx.doi.org/10.1109/ICT4DA59526.2023.10302229

[61] Zhu, X., Li, J., & Zhang, Y. (2023). Research on vehicle pricing factors of auto insurance based on machine learning. In 2023 8th International Conference on Communication, Image and Signal Processing (CCISP) (pp. 274–279). IEEE. https://dx.doi.org/10.1109/CCISP59915.2023.10355772

[62] So, B. (2023). CatBoost versus XGBoost and LightGBM: Developing enhanced predictive models for zero-inflated insurance claim data (Unpublished master's thesis). Towson University, Towson, MD.

[63] Manike, M., Singh, P., Madala, P. S., Varghese, S. A., & Sumalatha, S. (2021). Student placement chance prediction model using machine learning techniques [Paper presentation]. 2021 5th Conference on Information and Communication Technology (CICT), Kurnool, India. https://doi.org/10.1109/CICT53865.2020.9672372

[64] Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. arXiv. https://dx.doi.org/10.48550/arXiv.1604.06737

[65] Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826

[66] Earnix. (2024). AI & machine learning explainability in insurance. Retrieved from www.earnix.com/blog/ai-machine-learning-explainability-in-insurance/

[67] Richman, R., & Wüthrich, M. V. (2023). LocalGLMnet: Interpretable deep learning for tabular data. Scandinavian Actuarial Journal, 2023(1), 71–95. https://doi.org/10.1080/03461238.2022.2081816

[68] Vel, D. V. T., & Durgaraju, S. (2024). Explainable AI (XAI) for health insurance underwriting. International Journal of Communication Networks and Information Security, 15(1), 285–306. Retrieved from https://www.ijcnis.org/index.php/ijcnis/article/view/7252

[69] Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2023). Revisiting deep learning models for tabular data. arXiv. https://dx.doi.org/10.48550/arXiv.2106.11959

[70] McDonnell, K., Murphy, F., Sheehan, B., Masello, L., & Castignani, G. (2023). Deep learning in insurance: Accuracy and model interpretability using TabNet. Expert Systems with Applications, 217, Article 119543. https://doi.org/10.1016/j.eswa.2023.119543

[71] Thanapol, P., Lavangnananda, K., Bouvry, P., Pinel, F., & Leprévost, F. (2020). Reducing overfitting and improving generalization in training Convolutional Neural Network (CNN) under limited sample sizes in image recognition [Paper presentation]. 2020 5th International Conference on Information Technology (InCIT), Chonburi, Thailand. https://doi.org/10.1109/InCIT50588.2020.9310787

[72] Murad, N. Y., Ahmad, A., & Khan, S. (2024). Unraveling the black box: A review of explainable deep learning healthcare techniques. IEEE Access, 12, 66556–66568. https://dx.doi.org/10.1109/ACCESS.2024.3398203

[73] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? (arXiv:2207.08815). arXiv. https://doi.org/10.48550/arXiv.2207.08815

[74] Shwartz-Ziv, R., & Armon, A. (2021). Tabular data: Deep learning is not all you need. arXiv. https://dx.doi.org/10.48550/arXiv.2106.03253

[75] Cooper, A. (2022). Industry perspective: Tree-based models vs deep learning for tabular data. Retrieved from www.aidancooper.co.uk/tree-based-models-vs-deep-learning

[76] Aggarwal, C. C. (2018). Neural networks and deep learning: A textbook. Cham, Switzerland: Springer.

[77] Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of massive datasets (2nd ed.). Cambridge: Cambridge University Press.

[78] Hamilton, W. L. (2020). Graph representation learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 14(3), 1–159. https://doi.org/10.2200/S01045ED1V01Y202009AIM047

[79] Kotiyal, A., et al. (2024). Graph-based machine learning approaches for fraud detection in financial networks [Paper presentation]. 2024 7th International Conference on Contemporary Computing and Informatics (IC3I), Greater Noida, India. https://doi.org/10.1109/IC3I61595.2024.10828743

[80] Rakesh, M., & D, P. S. (2024). Enhanced medicare fraud detection using graph convolutional networks [Paper presentation]. 2024 4th International Conference on Artificial Intelligence and Signal Processing (AISP), Vijayawada, India. https://doi.org/10.1109/AISP61711.2024.10870744

[81] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods [Paper presentation]. AAAI/ACM Conference on AI, Ethics, and Society (AIES), New York, NY, United States. https://doi.org/10.1145/3375627.3375830.

[82] Hastie, T., Tibshirani, R., & Friedman, J. (2008). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York, NY: Springer.

[83] He, R., Li, Y., & Sun, C. (2025). Understanding software defect prediction through explainable neural additive models. IEEE Access, 13, 82860–82873. https://doi.org/10.1109/ACCESS.2025.3567140

[84] Walters, B., Jones, E., & Smith, J. (2022). Towards interpretable machine learning for clinical decision support. In 2022 International Joint Conference on Neural Networks (IJCNN) (pp. 1–8). IEEE. https://dx.doi.org/10.1109/IJCNN55064.2022.9892114

[85] Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a 'right to explanation'. AI Magazine, 38(3), 50–57. https://dx.doi.org/10.1609/aimag.v38i3.2741

[86] Monetary Authority of Singapore. (2018). Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of artificial intelligence and data analytics in Singapore's financial sector. Retrieved from www.mas.gov.sg/publications/monographs-or-information-papers/2018/feat

[87] South Korea. (2024). Basic Act on the Development of Artificial Intelligence and the Establishment of Foundation for Trustworthiness (AI Basic Act). Korean Official Gazette.

[88] Australia. (2024). Privacy Act 1988 (Cth). Federal Register of Legislation. Retrieved from www.legislation.gov.au/Details/C2024C00001

[89] Australia. Department of Industry, Science and Resources. (2019). Australia's artificial intelligence ethics principles. Retrieved from www.industry.gov.au/publications/australias-artificial-intelligence-ethics-principles/australias-ai-ethics-principles

[90] Ghana. (2012). Data Protection Act, 2012 (Act 843). Data Protection Commission. Retrieved from dataprotection.org.gh/wp-content/uploads/2023/12/Data-Protection-Act.pdf

[91] Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671–732. https://dx.doi.org/10.2139/ssrn.2477899

[92] Great Learning. (2024). Ensemble learning with stacking and blending | What is ensemble learning? Retrieved from www.mygreatlearning.com/blog/ensemble-learning/

[93] Brownlee, J. (2021). Blending ensemble machine learning with Python. Retrieved from machinelearningmastery.com/blending-ensemble-machine-learning-with-python/

[94] Chatzimparmpas, A., Bakoč, T., & Vukčević, M. (2021). Empirical study: Visual analytics for comparing stacking to blending ensemble learning. In 2021 23rd International Conference on Control Systems and Computer Science (CSCS) (pp. 1–8). IEEE. https://dx.doi.org/10.1109/CSCS52396.2021.00008

[95] Van Bekkum, M., de Boer, M., van Harmelen, F., & Meyer-Vitali, A. (2021). Modular design patterns for hybrid learning and reasoning systems. Applied Intelligence, 51, 6528–6546. https://dx.doi.org/10.1007/s10489-021-02394-3

[96] Darwiche, A. (2018). Human-level intelligence or animal-like abilities? Communications of the ACM, 61(10), 56–67. https://dx.doi.org/10.1145/3271625

[97] Bechler-Speicher, M., Globerson, A., & Gilad-Bachrach, R. (2024). TREE-G: Decision trees contesting graph neural networks (arXiv:2207.02760). arXiv. https://doi.org/10.48550/arXiv.2207.02760

[98] Good, J. H., Bouthillier, X., & Brooks, E. (2023). Feature learning for interpretable, performant decision trees. In NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems.

[99] Johansson, U., Sönströd, C., & Löfström, T. (2011). Trade-off between accuracy and interpretability for predictive in silico modeling. Future Medicinal Chemistry, 3(6), 647–663. https://dx.doi.org/10.4155/fmc.11.23

[100] Ahmed, U., Lin, J. C.-W., & Srivastava, G. (2024). Explainable AI-based innovative hybrid ensemble model for intrusion detection. Journal of Cloud Computing, 13, 150. https://dx.doi.org/10.1186/s13677-024-00712-x

[101] Lopardo, G., Precioso, F., & Garreau, D. (2024). Attention meets post-hoc interpretability: A mathematical perspective [Paper presentation]. 41st International Conference on Machine Learning (ICML), Vienna, Austria. https://openreview.net/forum?id=wnkC5T11Z9

[102] Mylonas, N., Mollas, I., & Tsoumakas, G. (2022). An attention matrix for every decision: Faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classification. arXiv. https://dx.doi.org/10.48550/arXiv.2209.10876

[103] Garcez, A. d., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. arXiv. https://dx.doi.org/10.48550/arXiv.1905.06088