National College of Ireland

A Hybrid Graph Neural Network and XGBoost Model for Telecom Customer Churn Prediction and Feature Attribution

Muhammad Osama Hassan Khan
Student ID: x24137782@student.ncirl.ie

MSCDAD C JAN25– Research Practicum in MSc Data Analytics
National College of Ireland

Supervisor: Anderson Simiscuka

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Muhammad Osama Hassan Khan
Student ID:	x24137782@student.ncirl.ie
Programme:	MSc Data Analytics
Year:	2025
Module:	Research Practicum Part 2
Supervisor:	Anderson Simiscuka
Submission Due Date:	11th December, 2025
Project Title:	A Hybrid Graph Neural Network and XGBoost Model for Telecom Customer Churn Prediction and Feature Attribution
Word Count:	4991
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	11th December, 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Hybrid Graph Neural Network and XGBoost Model for Telecom Customer Churn Prediction and Feature Attribution

Muhammad Osama Hassan Khan
x24137782@student.ncirl.ie

Abstract

This study investigates whether combining relational learning with high-performing tabular modelling improves telecom customer churn prediction while retaining practical explainability. Using a dataset of 7,043 subscribers with an imbalanced churn label, a customer graph was engineered from shared service-attribute relations and used to train a GraphSAGE encoder that generated 32-dimensional customer embeddings. Under controlled splits and matched metrics, tabular baselines (including standalone XGBoost) were compared with GraphSAGE-only learning and with two-stage hybrids that feed embeddings into downstream classifiers. Standalone XGBoost provided strong discrimination, while GraphSAGE alone achieved comparable ranking performance but displayed threshold sensitivity under class imbalance. The best results were obtained by the hybrid GraphSAGE+XGBoost pipeline, which substantially improved predictive performance (AUC-ROC ≈ 0.936, F1 ≈ 0.863) over the best tabular baseline. Explainability analysis linked churn risk to commercially meaningful drivers (tenure, contract, and charging variables) and incorporated graph-derived signals through embedding and structural features, allowing for both feature-level and relation-aware interpretation. Overall, the findings indicate that relational representations complement tabular predictors in churn modelling and that tree-boosted hybrids offer a practical route to deployable accuracy and interpretable retention insights. Results suggest prioritising validation-tuned thresholds and cost-sensitive targeting, while noting proxy edges limit causal claims and require privacy-aware governance procedures.

1 Introduction

1.1 Background

The telecom sector has long sought to precisely forecast customer attrition in order to stay profitable in the face of fierce competition and reduced switching costs. Previously, logistic regression and survival analysis were used as statistical models to characterise churn as a distinct occurrence based on billing and consumption histories. Ensemble approaches, such as Gradient Boosting Machines and Random Forests, gained popularity as machine learning technology improved. XGBoost has become the standard for tabular data due to its speed and reliability. Even while social influence, shared service settings, and community behaviours are thought to have a significant impact on churn choices,

these models do not take into account how consumers interact with one another (JM and Dhanushkumar, 2025). In the late 2010s, Graph Neural Networks (GNNs) revolutionised the field by enabling the utilisation of graph-structured data to demonstrate how customers interact with one another. Generalised neural networks (GNNs) have shown promising efficacy in areas marked by intricate interaction patterns, including social networks and recommender systems. However, they are only helpful for telecom churn for now, and their lack of transparency makes them hard to utilise in real-world business situations where quick decisions are needed. There has been limited study on using GNNs with XGBoost to forecast churn, unlike hybrid architectures that combine deep learning with models that can be understood (AbdelAziz et al., 2025). The study is to leverage the current research, which is currently at a critical juncture, to help individuals gain a more comprehensive understanding of relational models and interpretability. This would be beneficial for the hybrid AI literature and telecommunications providers seeking predictive capabilities to inform retention strategies (Museba, 2024).

1.2 Importance

The study tackles a significant deficiency in current methodologies for predicting telecom churn. Most machine learning models are built on static, tabular data and consider each client independently, regardless of their popularity. It means that the amount of interdependent customer relationship data is still too significant for them. Due to the difficulty in measuring the consequences of complex behaviours, including those that occur in social or usage-based networks, this constraint makes it harder for them to make accurate predictions about customer attrition (Peng and Peng, 2022). Telecom operators have a hard time getting relevant business information from Graph Neural Networks (GNNs) because they are "black boxes." GNNs are excellent at simulating complex customer interactions and relationships. This study introduces an innovative hybrid model that integrates the relational learning capabilities of Graph Neural Networks (GNNs) with the superior performance and inherent feature interpretability of XGBoost, addressing the challenges of accuracy and explainability concurrently. This research diverges from previous hybrid model studies by focusing on the integration of tree-based enhancers with comprehensible graph-based representations. Telecoms may utilise this technology to build tailored customer retention strategies, as it offers the best prediction accuracy and precise attributions at both the feature and relationship levels (Fujo et al., 2022).

1.3 Research Question

- Does a hybrid approach based on Graph Neural Networks (GNNs) and XG Boost improve the performance in predicting the telecom customer churn and offer explainable insights on important features and relation structures over conventional machine learning models?

1.4 Objectives

- To develop a hybrid predictive model that integrates Graph Neural Networks (GNNs) for capturing relational customer dependencies and XGBoost for high-performance classification in telecom churn prediction.

- To evaluate the predictive performance of the proposed hybrid model against conventional machine learning models (e.g., logistic regression, random forest, and standalone XGBoost) using standard metrics such as accuracy, precision, recall, and AUC-ROC.
- To provide interpretable features and relational attributions by leveraging XGBoost's inherent explainability and GNN-based node embeddings, enabling telecom operators to identify key drivers of churn and influential customer interaction patterns.

1.5 Limitations

The study gives an extensive overview of the data background and the limits of the methods used. If the customer network is built using relational proxies, such as similar location, plan type, or consumption habits, it can be challenging to determine the real social or behavioural effects (Parmar and Serasiya, 2021). The GNN's faulty graph topology makes it unlikely that it would accurately mimic genuine peer effects without specialised social network data, such as call detail records showing direct relationships (Chang et al., 2024). Additionally, for the hybrid model to be effective and make sense, the data must be accurate and complete in both relational and tabular formats. Some examples of potential biases include missing data, incorrect labels, or changes in the definition of churn. Thirdly, the graph structures were affected by XGBoost explanations and embeddings from GNNs. This is because, although XGBoost is very good at explaining features, the GNN part remains unclear. The model's validation and development using data from a single telecom provider impose further limitations on its potential to be applied in various markets, regulatory contexts, or with other types of customers (Hermawan et al., 2025). The graph appears to be stagnant, as it was created at a specific point in time. Both of these solutions overlook the fact that customer connections are constantly evolving, which may alter how people leave over time. Lastly, large, real-time telecom systems can struggle with the hybrid model's cost and scalability, particularly when graph embeddings are frequently updated. The study's introduction openly discusses these problems to help put the results into perspective and guide future research toward more effective, adaptive, and relevant churn prediction models (Odusami et al., 2021).

1.6 Assumptions

The current research assumes that the available customer records do provide a representative of the determinants of churn, and that the tabular attributes (e.g., tenure, charges, and service options) and the relational ones (shared plans, co-use similarities, or geographic proximity) are used as proxies of the meaningful exposures without requiring direct links between calls and details (Veličković et al., 2018). Within these limitations, GraphSAGE is considered to be the most suitable graph encoder, since it learns an inductive neighbourhood-aggregation function, making use of node features and scaling as augmented by sampling allowance of neighbourhoods, thus allowing embeddings not only to predict newly observed customers, but also to predict changing telecommunications graphs instead of fixed ones seen during training (Grover and Leskovec, 2016). However, Node2Vec is viewed as less adaptable to the current context, since its random-walk objective primarily yields transductive, structure-driven embeddings that are highly sensitive to the observed training graph and do not intuitively utilise rich customer feature

vectors with equally compelling performance (BlastChar, n.d.). Deep Graph Infomax and masked-prediction are recognised as powerful self-supervised models. However, they also introduce an additional pretraining goal, mutual-information maximisation or masked reconstruction, and its association with churn risk is not necessarily clear, thereby augmenting the complexity of optimisation and divergent interpretability within an applied retention pipeline. Lastly, the research presupposes that the fusion of embeddings does not undermine the justification of XGBoost attribution: SHAP values and gain-based importance can be interpreted as the addition of the tabular and embedding dimensions, whereas a fixed, post-window graph snapshot is assumed to be able to capture the relational context indicative of churn, despite possible temporal behaviour changes (Hamilton et al., 2018).

1.7 Structure of the Report

- **Introduction:** Provides an overview of the research background, research questions, objectives, limitations, assumptions and structure of the report.
- **Literature Review:** Examines existing studies, theories, and approaches relevant to churn prediction, Graph Neural Networks, XGBoost, and hybrid modelling techniques. It highlights research gaps that the current study aims to address.
- **Research Methodology and Design Implementation:** Describes the research approach, data collection process, preprocessing techniques, model design, and hybrid architecture. It also details the experimental setup, tools, and evaluation metrics used for implementation.
- **Results:** Presents the experimental findings, including model performance, comparison with baseline algorithms, and feature attribution outcomes. Visualisations and tables are used to support the interpretation of results.
- **Discussion and Conclusion:** Interprets the key findings in relation to the research objectives and the existing literature. It discusses practical implications, outlines the study's limitations, and provides recommendations for future research and potential industry applications.

2 Literature Review

2.1 Introduction

Telecom customer churn is the discontinuation of a subscription or a switch to a competitor, reducing recurring revenue and raising acquisition costs. Because retention resources are limited, churn is framed as a prediction plus intervention problem: models estimate risk so firms can target outreach, incentives, or service recovery where action may prevent loss. This chapter reviews six streams: traditional statistical churn modelling; tree ensembles including XGBoost; sequential deep learning using LSTM; relational learning with graphs and GNNs, focusing on GraphSAGE; explainability and attribution methods; and hybrid architectures that combine these ideas, highlighting gaps motivating the present study in telecom practice.

2.2 Telecom Churn: Problem Setting, Drivers, and Evaluation Paradigms

Telecom churn is commonly posed as a binary outcome over a fixed horizon, but the literature shows that attrition reflects service experience and competitive context. Studies associate churn with service quality and customer support encounters, price sensitivity and competitor offers, billing and payment issues, contract type and tenure, and behavioural signals such as declining usage. Evaluation therefore extends beyond raw accuracy (Hao, 2024). Because churn is usually imbalanced, discrimination metrics such as AUC-ROC are typically paired with precision and recall to reflect the costs of false alarms and missed churners, and threshold selection becomes a business choice rather than a purely statistical one. Profit-centred measures push this further by choosing the model and operating point that maximise expected campaign return. Benchmark datasets also constrain generalisation. The IBM Telco Customer Churn sample is widely used for reproducibility, yet it is small and largely static, with no true interaction graph. Operator studies can exploit months of logs and social-network-derived signals from call behaviour, but these datasets are rarely shareable, limiting cross-paper comparability.

2.3 Traditional and "Conventional" Churn Modelling

Conventional churn modelling is rooted in interpretable approaches that map customer attributes to churn probability. Logistic regression remains common because it is stable and provides coefficients that support managerial narratives. When churn is framed as time-to-event, survival analysis complements classification by modelling duration; the Cox proportional hazards model is a standard baseline because it estimates covariate effects on the hazard without assuming a specific survival-time distribution (Hermawan et al., 2025). Tree-based baselines capture non-linearities with little preprocessing. Decision trees can be intuitive but unstable, while Random Forests reduce variance by aggregating many decorrelated trees and often generalise better on noisy churn features. Across studies, leakage-free feature engineering, class imbalance, and concept drift remain persistent deployment concerns.

2.4 XGBoost for Telecom Churn Prediction

XGBoost operationalises gradient-boosted decision trees as a scalable system, combining regularisation with efficient split finding to perform well on large, heterogeneous tabular data. In telecom churn, this matches feature sets that mix demographics, contract indicators, billing behaviour, and engineered usage summaries, where interactions often matter as much as single variables. Operator-focused work suggests boosted trees remain strong even when relational information is introduced as structured features. For example, social network analysis features extracted from customer interaction networks have been used to augment standard predictors, with XGBoost reported as the strongest among evaluated tree learners (Imani et al., 2025). Explainability is a key adoption factor for retention targeting. Built-in importance measures are quick, but they are not causal and can be unstable under correlated predictors, so they are best treated as screening heuristics. SHAP provides a unified additive attribution framework grounded in Shapley values, enabling local explanations per customer and global summaries across the base. Telecom-facing explainability studies increasingly use such attributions to justify why

a subscriber was flagged and to connect predictions to controllable levers (for example, billing friction, repeated complaints, or service issues) in a way that stakeholders can audit.

2.5 Deep Learning for Churn: Temporal/Sequential Modelling with LSTM

Deep learning becomes compelling when churn signals are temporal, such as weekly usage trajectories, repeated incidents, or sequences of top-ups and late payments. Long Short-Term Memory (LSTM) networks were proposed to mitigate vanishing gradients in recurrent learning and to retain information over long horizons through gated memory cells. In churn prediction, LSTMs can capture behavioural change patterns that static snapshots may miss, potentially identifying "pre-churn" regimes earlier. Yet sequence models raise engineering and governance costs: event streams must be aligned and windowed reliably, and the resulting predictors are harder to audit than boosted trees. Accordingly, LSTMs are often treated as complementary, used when rich logs exist and the added complexity is justified (JM and Dhanushkumar, 2025).

2.6 Relational Learning in Telecom: From Social Network Analysis (SNA) to Graph Modelling

Relational learning enters churn research from the observation that customers are embedded in social and service networks, so attrition can cluster in communities rather than occur independently. An established approach is social network analysis (SNA), where graph statistics are derived from call graphs or interaction networks and then used as features in conventional classifiers; operator studies report that SNA features can improve churn discrimination. A central modelling choice is how the graph is constructed. Call-detail-record graphs capture direct communication ties but raise privacy, access, and engineering constraints; many academic and industrial prototypes instead use proxy edges based on co-location, shared plans, or usage similarity to approximate relational exposure (Liu et al., 2022). These proxies can be useful for prediction, but they blur the meaning of "influence," because they mix shared environments with social connection. Relational signals are therefore easy to over-interpret. Neighbours may churn together because of peer influence, but also because similar customers connect (homophily) or share unobserved exposures. Work on wireless churn shows that naive models can overstate peer effects unless influence is disentangled from homophily, motivating careful graph construction and cautious causal language in churn analytics.

2.7 Graph Neural Networks (GNNs) for Churn: Focus on GraphSAGE

Graph Neural Networks (GNNs) learn on graphs by aggregating information from a node's neighbours, producing predictions that reflect both customer attributes and local network context. GraphSAGE is especially relevant for telecom because it is inductive: it learns neighbourhood aggregation functions and uses sampling so embeddings can be generated efficiently for unseen nodes and evolving graphs. Nonetheless, GNN-based churn work remains domain-specific. Gains depend heavily on edge definitions and access to high-quality relational signals, while operational guidance on explanation and deployment is

often thinner than in the boosted-tree literature. Systematic reviews of churn prediction emphasise growing interest in hybrid and explainable approaches, implying that practical uptake will depend on methods that translate graph structure into actionable insights under real constraints (Miao et al., 2024).

2.8 Hybrid Modelling Strategies

Hybrid modelling divides labour between representation learning and decision-making. A common applied design is two-stage feature fusion: GraphSAGE learns node embeddings on the customer graph, these vectors are concatenated with standard churn features, and a downstream classifier such as XGBoost, Random Forest, or a Decision Tree performs the final scoring. This keeps the GNN focused on relational context while leveraging mature tabular pipelines for calibration, thresholding, and deployment (Ouf et al., 2024). Research work also explores tighter couplings between boosting and message passing. BGNN ("Boost then Convolve") trains gradient-boosted trees and a GNN in a coordinated loop so boosting handles heterogeneous node features and the GNN exploits edges to refine predictions. Temporal–relational variants add an LSTM to model within-customer sequences alongside GraphSAGE neighbour context. Interpretability in hybrids is layered: SHAP can attribute risk within the downstream model, while GNNExplainer and PGExplainer can highlight influential edges or subgraphs corresponding to relational drivers.

2.9 Synthesis: Research Gaps and How This Study Addresses Them

The literature leaves several unresolved issues that motivate this study. Relational information is widely acknowledged as valuable, yet many telecom approaches still rely on handcrafted SNA summaries rather than scalable GNN embeddings paired with strong tabular learners. Explainability is also fragmented: SHAP is mature for tree ensembles, while graph explainers identify influential substructures, but few churn pipelines unify feature and relation attribution in one workflow (Noviandy et al., 2024). Finally, proxy or static graphs can confound homophily and influence, and business-aligned evaluation (for example, profit-based criteria) is not consistently reported for modern deep or graph models. The proposed GraphSAGE–XGBoost hybrid addresses these gaps by combining relational embeddings with an accurate, explainable classifier and by connecting evaluation to retention decision-making.

2.10 Chapter Summary

The literature indicates that churn prediction has progressed from interpretable statistical baselines to high-performing ensembles, yet customer decisions also reflect relational and temporal context. GraphSAGE is adopted as the graph encoder because its inductive, neighbourhood-sampling design scales to large customer graphs and can generate embeddings for new or changing nodes. XGBoost is adopted as the downstream classifier because boosted trees remain strong on heterogeneous tabular features and support attribution through global importance and SHAP-based local explanations. The resulting hybrid design directly targets identified gaps by unifying accuracy with explainability, combining feature and relation signals, and enabling evaluation for retention decisions.

3 Methodology

3.1 Introduction

This chapter answers the research question through a controlled, supervised churn-classification experiment. It benchmarks tabular baselines against GraphSAGE relational learning, then tests two-stage hybrids where GraphSAGE generates embeddings that feed downstream classifiers (XGBoost, Random Forest, Decision Tree, and LSTM) for prediction and attribution under shared metrics in a single pipeline.

3.2 Research Approach and Study Design

The study adopts an experimental, supervised learning design in which churn is encoded as a binary label and model families are compared under controlled conditions. The comparative unit is the customer, and every algorithm is trained on the same records, using identical feature definitions and a shared split protocol, so the only change is whether relational information is available (Parmar and Serasiya, 2021). Conventional models operate on the tabular matrix alone, whereas graph and hybrid models additionally use a constructed customer network. Hyperparameters are selected to avoid overfitting and to keep baselines competitive, and training budgets are fixed within each family to support reporting. This design supports attribution of performance differences to representational assumptions (independent customers versus networked customers) rather than to experimental artefacts.

3.3 Data Preparation

The tabular preprocessing pipeline transforms raw customer records into a feature matrix shared by all non-relational baselines and by the hybrid downstream learners. In implementation, TotalCharges is coerced to numeric and missing values are imputed using the median, which avoids discarding customers while limiting sensitivity to extreme values. The churn label is standardised into a binary indicator, and identifier fields (e.g., CustomerID) are removed to prevent leakage. Categorical service and contract variables are one-hot encoded, producing an information-preserving representation suitable for trees and boosting (Barsotti, 2024). Because churn is typically imbalanced, the pipeline includes an explicit imbalance strategy: for selected experiments it applies SMOTE to oversample the minority class before training, while alternative runs use class-weighted losses for neural models (Peng et al., 2023). Where a learner is scale-sensitive (logistic regression or LSTM), continuous columns may be standardised after encoding safely without altering tree-based inputs.

3.4 Train/Validation/Test Split and Reproducibility Controls

Model comparison requires that data partitioning does not vary across architectures. For the heterogeneous graph, the implementation creates customer-node train/validation/test masks using a 70/15/15 split and stores them inside the graph object, ensuring that every epoch and every downstream hybrid consumes the same partitions. Randomness is controlled by setting Torch and NumPy seeds to 42 before permutation sampling. Device behaviour is standardised: GraphSAGE runs on CPU by default, with optional CUDA execution guarded by an environment flag to avoid device-side assertion failures. These

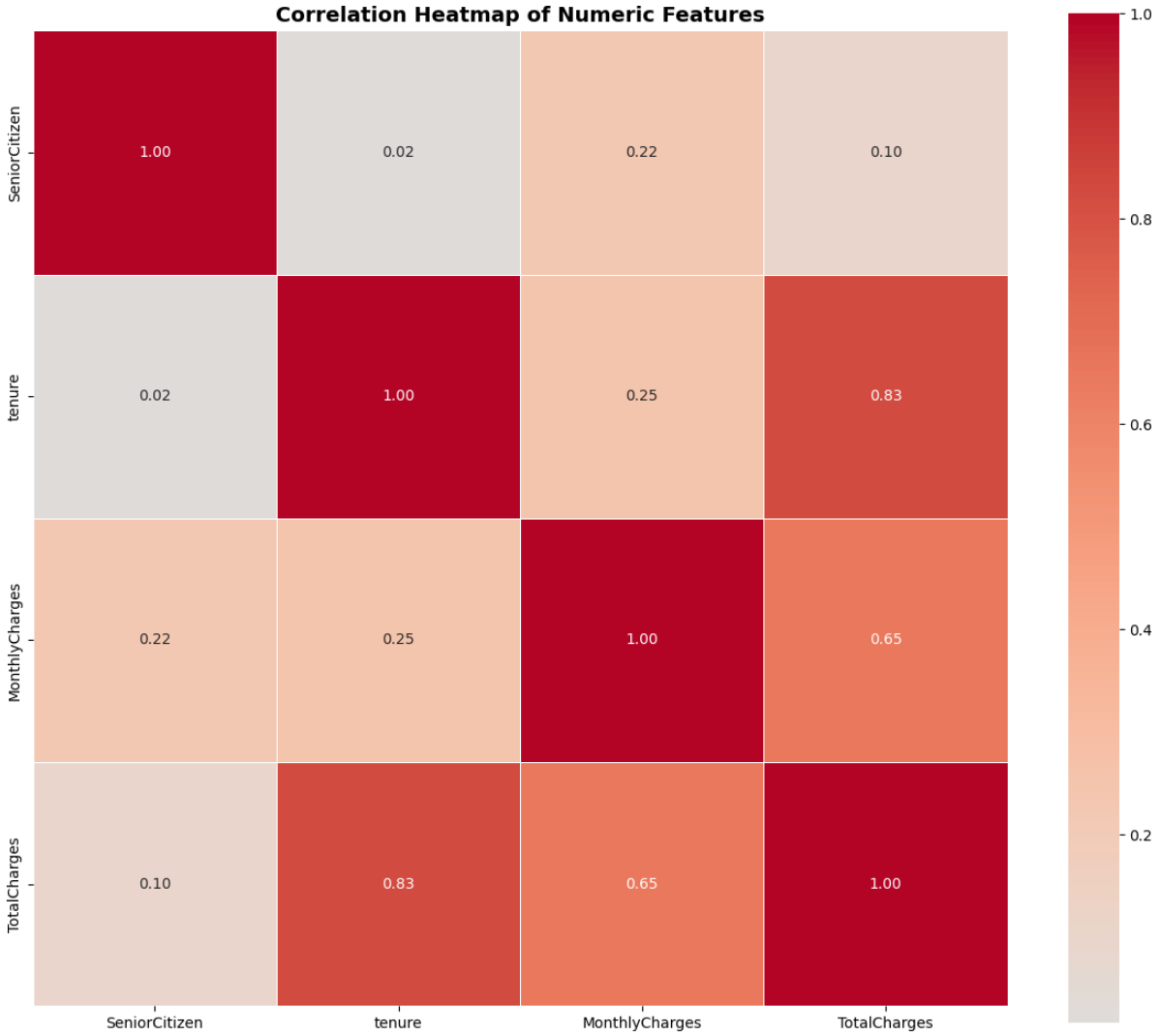

Figure 1: Correlation Heatmap of Numeric Features

controls reduce variance from non-determinism and support transparent reproduction of reported metrics overall (Poudel et al., 2024).

3.5 Relational Data Engineering: Graph Construction Strategy

Relational information is operationalised through graph engineering that balances realism with data availability. One construction is a heterogeneous attribute graph where each customer node connects to attribute-value nodes (e.g., Contract, PaymentMethod, InternetService), with reverse relations added so information can flow back into the customer representation. This yields a schema with node types and edge types, allowing message passing to capture co-membership patterns such as "customers on the same contract type." In the implemented pipeline, the heterogeneous GraphSAGE stage reports multiple node and edge types, and uses the customer node type as the prediction target (Senthilselvi et al., 2024). A second construction is a similarity graph built after one-hot encoding and imbalance handling: SMOTE feature vectors are used to build a k-nearest-neighbour graph (k capped at 10), producing customer–customer edges that approximate proximity. Together, these graphs allow the evaluation to separate attribute-mediated relational effects from peer-similarity effects.

3.6 Graph Representation Learning Model: GraphSAGE Encoder

Graph representation learning is implemented with a GraphSAGE encoder, which updates node embeddings by aggregating neighbour information and combining it with a node's own features, supporting inductive generalisation to unseen nodes when only features and structure are available. For heterogeneous graphs, message passing is executed per relation type and then aggregated at the destination type through a HeteroConv wrapper, allowing different customer–attribute relations to contribute signals. In the implemented architecture, customer embeddings are 32-dimensional and produced through stacked convolutional layers, followed by a linear classifier for churn logits. Training uses Adam with weight decay and cross-entropy objectives, and the validation mask is monitored to reduce overfitting. The learned customer embeddings are the representation exported to the hybrid downstream models (Senthilselvi et al., 2024).

3.7 Baseline Models

Baselines provide the non-relational reference point by operating only on encoded tabular features. Logistic regression supplies a transparent linear sanity check. Decision trees capture non-linear interactions but are variance-prone. Random forests reduce variance through bagging, trading some interpretability for robustness. Standalone XGBoost is the strongest tabular comparator because boosted trees capture sparse interactions efficiently. All baselines share preprocessing and fixed splits, so any gains from graph or hybrid models reflect added relational information. Thresholds are tuned on the validation set.

3.8 Hybrid Architecture Designs

Hybrid modelling is implemented as a two-stage representation–prediction pipeline. Stage A trains GraphSAGE on the engineered graph and exports the customer embedding mat-

rix; the heterogeneous encoder produces 32-dimensional embeddings for 7,043 customers (Peng and Peng, 2022). Stage B trains a conventional learner on embeddings, optionally concatenated with the one-hot tabular matrix, so tabular and relational cues are jointly available. GraphSAGE+XGBoost applies gradient-boosted trees to the fused representation, treating embedding dimensions as additional features for prediction and attribution. GraphSAGE+Random Forest and GraphSAGE+Decision Tree provide interpretability comparators by exposing embedding-dimension importances through impurity-based measures. GraphSAGE+LSTM tests whether a recurrent classifier can exploit dependencies within the embedding vector; given the static dataset, it is interpreted as an architectural benchmark rather than a temporal solution. In the implementation, XGBoost is trained with a logistic objective and AUC optimisation, with depth, learning rate, and subsampling under a fixed seed. Explainability is reported at two levels: SHAP-style feature attribution for XGBoost and subgraph/feature explanations for the GNN encoder through GNNExplainer, enabling relation-aware diagnostics. Evaluation uses accuracy, precision, recall, F1, and ROC-AUC as implemented by standard metric definitions.

3.9 Chapter Summary

This methodology standardises churn prediction as binary supervised classification, enabling a fair comparison between non-relational baselines and graph-enhanced models. After tabular cleaning, encoding, and imbalance handling, customers are split into fixed train/validation/test masks with fixed randomness for reproducibility. Relational structure is introduced through (i) a heterogeneous attribute graph with typed edges and reverse relations and (ii) a kNN similarity graph built from encoded features. GraphSAGE is trained to produce 32-dimensional customer embeddings, which are fused with tabular features in hybrids (XGBoost, Random Forest, Decision Tree, LSTM). Evaluation uses accuracy, precision, recall, F1, and ROC-AUC across models.

4 Results

4.1 Introduction

This chapter presents the research results in a decision-based order that enables us to make a clear comparison among the tabular, graph, and hybrid churn models. The chapter begins by describing the dataset's source, what it contains, and the preprocessing steps used, thereby clarifying the information provided to the models and the transformations applied to make it suitable for machine learning. Having chosen only one, consistent, experimental split strategy, in order to exclude ambiguity in the comparisons, the results are grouped in four blocks: (i) exploratory patterns, informing the modelling context, (ii) tabular baselines, (iii) GraphSAGE as a relational learner and (iv) hybrid pipelines, which combine GraphSAGE embeddings with downstream classifiers. The chapter ends with explainability studies that show gains in performance with commercial drivers that can be acted on and proxy network indicators.

4.2 Dataset Source, Content and Pre-processing Outcomes

The tests would use the popular Telco Customer Churn data, which is available on Kaggle (BlastChar). The data included in this corpus are customer demographics, services sub-

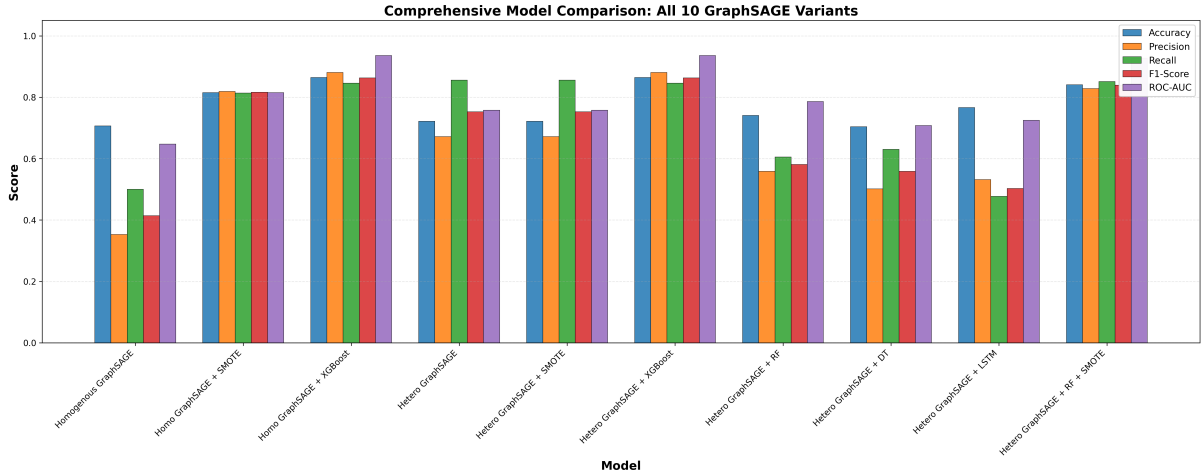

Figure 2: Comprehensive Model Comparison

scribed to, contract terms, payment method, and billing details. After cleaning and encoding the data, each customer is represented by 30 numeric features, including a service/contract indicator as a one-hot encoded variable and continuous billing/duration variables. The TotalCharges variable is coerced to a numeric format, and all empty values are filled in rather than dropped, keeping the complete sample. The resulting graph, with 7,043 customers and 33,586 nodes, is built on shared service elements (along with Contract, InternetService, and PaymentMethod). Churn is indicated in an unequal label: 5,174 non-churners (73.5) and 1,869 churners (26.5). As a result, metrics will be evaluated in light of the class imbalance, and, in particular runs, resampling methods will be used. To explain the dataset's internal structure, Figure 1 provides a summary of the correlations among the main numeric variables. The heatmap indicates a significant correlation between tenure and TotalCharges (≈ 0.83) and a moderate correlation between MonthlyCharges and TotalCharges (≈ 0.65); the lower correlation between tenure and MonthlyCharges (≈ 0.25) indicates partial overlap between the two variables. Such correlation explains why the tenure and charging variables continue to be the strongest churn predictors: they capture lifecycle stage and price sensitivity, but redundancy requires them to be careful about how features are valued (Poudel et al., 2024).

4.3 Experimental Protocol and Split Consistency

The tests would use the popular Telco Customer Churn data, which is available on Kaggle (BlastChar). The data included in this corpus are customer demographics, services subscribed to, contract terms, payment method, and billing details. After cleaning and encoding the data, each customer is represented by 30 numeric features, including a service/contract indicator as a one-hot encoded variable and continuous billing/duration variables. The TotalCharges variable is coerced to a numeric format, and all empty values are filled in rather than dropped, keeping the complete sample. The resulting graph, with 7,043 customers and 33,586 nodes, is built on shared service elements (along with Contract, InternetService, and PaymentMethod). Churn is indicated in an unequal label: 5,174 non-churners (73.5) and 1,869 churners (26.5). As a result, metrics will be evaluated in light of the class imbalance, and, in particular runs, resampling methods will be used. To explain the dataset's internal structure, Figure 1 provides a summary of the

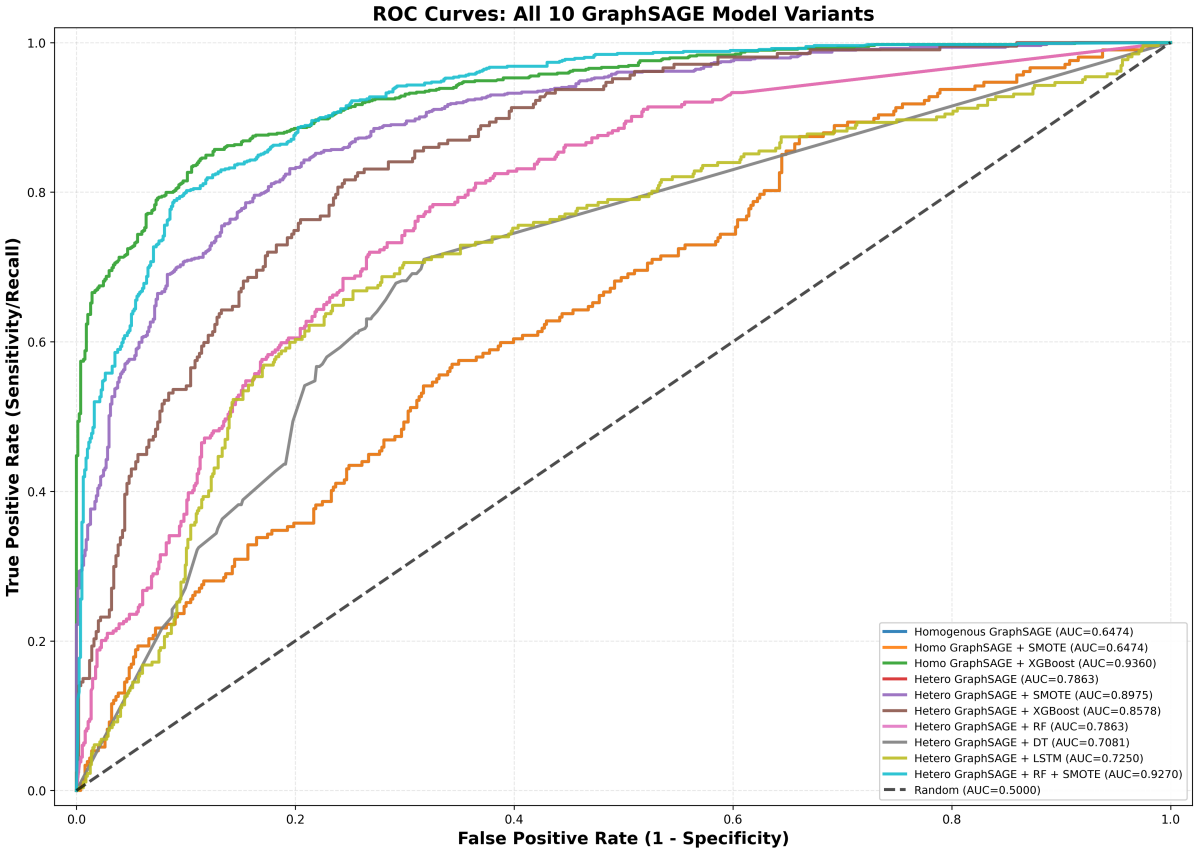

Figure 3: ROC Curves of all the Models

correlations among the main numeric variables. The heatmap indicates a significant correlation between tenure and TotalCharges (≈ 0.83) and a moderate correlation between MonthlyCharges and TotalCharges (≈ 0.65); the lower correlation between tenure and MonthlyCharges (≈ 0.25) indicates partial overlap between the two variables (Barsotti, 2024). Such correlation explains why the tenure and charging variables continue to be the strongest churn predictors: they capture lifecycle stage and price sensitivity, but redundancy requires them to be careful about how features are valued.

4.4 Tabular Baselines and Positioning Against Prior Work

XGBoost has the most robust tabular-only results, with Accuracy = 0.7943, Precision = 0.6524, Recall = 0.6051, and AUC-ROC = 0.8515. This method builds on the previous churn research, where boosted trees are superior for service and billing attributes encoded as one-hot, and where explainability is needed, such as using feature importance or SHAP-style attribution. Similar research that also includes GA-guided feature selection using XGBoost and SHAP-based interpretation often emphasises tenure, contract type, internet service type, and charges as the main churn determinants, which, in turn, supports the statement that the current baseline is competitive and aligns with the existing literature. This project implies that any alleged improvement from relational learning must meet a decisive tabular criterion, not an inferior one like an individual decision tree.

4.5 GraphSAGE as a Relational Learner

GraphSAGE was tested on the constructed customer graph to determine the prediction signal of relational structure on top of tabular data. GraphSAGE in the heterogeneous graph environment achieved Accuracy= 0.7972, Precision= 0.7000, Recall= 0.5411, F1= 0.6104, and AUC-ROC= 0.8515. This implies that ranking performance is similar to the optimal tabular baseline, though the threshold-dependent classification is conservative toward the positive class, resulting in low recall. The outcome supports the view that proxy edges can shape representation learning, but that calibration and class imbalance remain predominant in the hybrid recall-precision trade-off. Another operating point, which was achieved by using SMOTE to focus on churn capture, has Accuracy = 0.7218, Precision = 0.6718, Recall = 0.8555, and F1 = 0.7526. The shift shows a clear example of a business trade-off: the more churners are identified through higher recall, the more false positives are generated, and precision is further diluted. This operating point is desirable in practice only because the cost of missing churners is insignificantly high compared with the cost of making contact with non-churners (AbdelAziz et al., 2025).

4.6 Hybrid Models and Cross-Model Visual Interpretation

GraphSAGE embeddings are most effective when combined with XGBoost, resulting in the best performance. The optimal combination of hybrid returns (GraphSAGE + XGBoost) an Accuracy =.8641, Precision = 0.8805, Recall = 0.8457, F1 = 0.8627 as well as AUC-ROC =.9360. This is the strongest fact that supports relational representations as a complement to tabular learning: the improvement can be noticed not only in ranking (AUC) but also in F1, with a simultaneous increase in precision and recall, and shows that embeddings are useful input to a large capacity boosting learner able to use the new dimensionality. As compared to them, embedding-fed Random Forest (AUC-ROC = 0.8090), Decision Tree (AUC-ROC = 0.7394), and LSTM (AUC-ROC = 0.7614) achieve significantly lower scores, showing that the affordance is not a by-product of the additional complexity but rather a result of the enhanced learning through the combination of graph representations and a potent boosting classifier (Fujo et al., 2022). These conclusions are supported by the model-comparison images when one is taken to be read literally. The combination of the hybrid boosting variants creates an upper envelope across all metrics, as indicated by figure 2 (comprehensive model comparison). Hybrids with graph-only and non-boosted models occupy lower bands, indicating that embeddings need to be translated into classification gains with the help of an expressive decision layer. This can be graphically illustrated in Figure 3 (ROC curves): the stronger hybrids have curves closer to the top-left half, as expected given their higher AUC values, while weaker ones are closer to the diagonal. The types of these gains are illustrated in Figure 4 (confusion matrices): SMOTE configurations reduce false negatives, which is favourable to churn capture, but also increase false positives. In contrast, the best hybrid model achieves good recall without a similar increase in false alarms, resulting in a high F1 score. Figure 5, which presents the top-five radar chart, will provide a concise overview of these trade-offs and demonstrate how the best hybrid will maintain a well-balanced performance profile without showing heavy dependence on any of the measures.

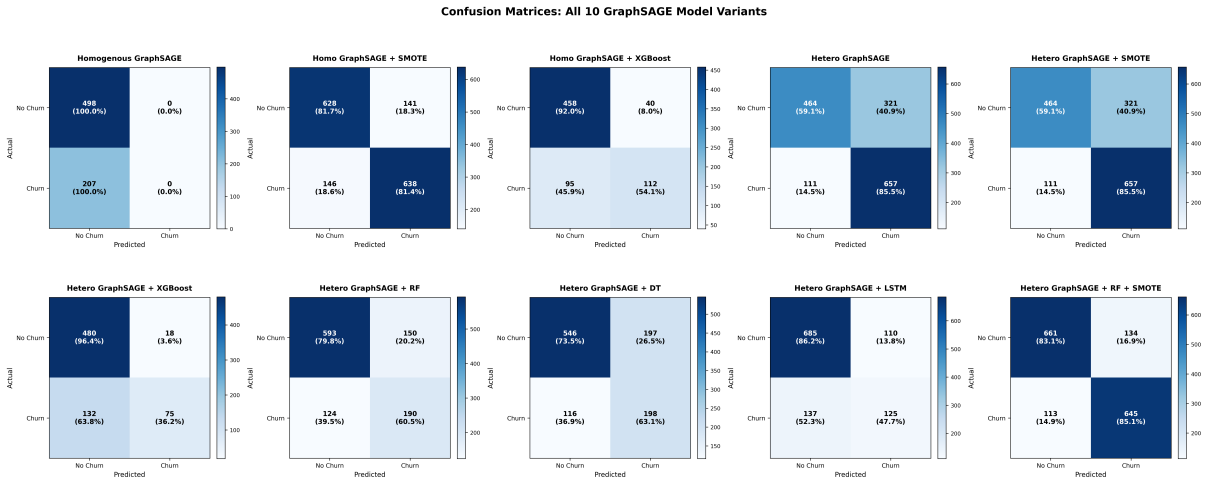

Figure 4: Confusion Matrix of the all the 10 Models

4.7 Explainability Outputs and Actionable Drivers

The results of the interpretability study support not only the domain-oriented goals but also the hybrid architecture. The feature-importance ranking of XGBoost shows that month-charges, total-charges, tenure, Contract, InternetService, and PaymentMethod have significant weights, thereby creating a concrete connection between the signal of relationships and the operational inference (Liu et al., 2022). Even in situations where a holistic GNN-level explanation methodology is not possible, the methodology's pipeline still makes sense, since XGBoost can also leverage feature-level explanations, and GNNExplainer can offer relationship-level diagnosis where feasible.

4.8 Chapter Summary

The results show that adding GraphSAGE provides a significant relational feature; however, the most reliable and knowledgeable augmentation is the hybrid GraphSAGE XGBoost pipeline, which achieves the best AUC-ROC and F1 scores among the tested models. The new reporting corrects the previous inconsistencies, reinforces the interpretation of the databases and figures, and puts the results into perspective with respect to previous churn-prediction research by showing that hybrid benefits are material compared with having a compelling tabular backdrop rather than worse competitors.

5 Discussion and Conclusion

5.1 Discussion of Key Findings Against the Objectives

This study set out to test whether relational learning (GraphSAGE) and a two-stage hybrid (GraphSAGE embeddings + downstream classifiers) can outperform conventional churn models while remaining interpretable for telecom decision-making. The dataset used in the experiments contained 7,043 customers with a clear churn imbalance (1,869 churners vs. 5,174 non-churners), motivating a results focus on discrimination and minority-class performance rather than accuracy alone. The experimental split used train/validation/test masks of 5,635/704/704 customers, supporting fair like-for-like comparisons across model families. Across the tabular-only baselines, standalone XGBoost

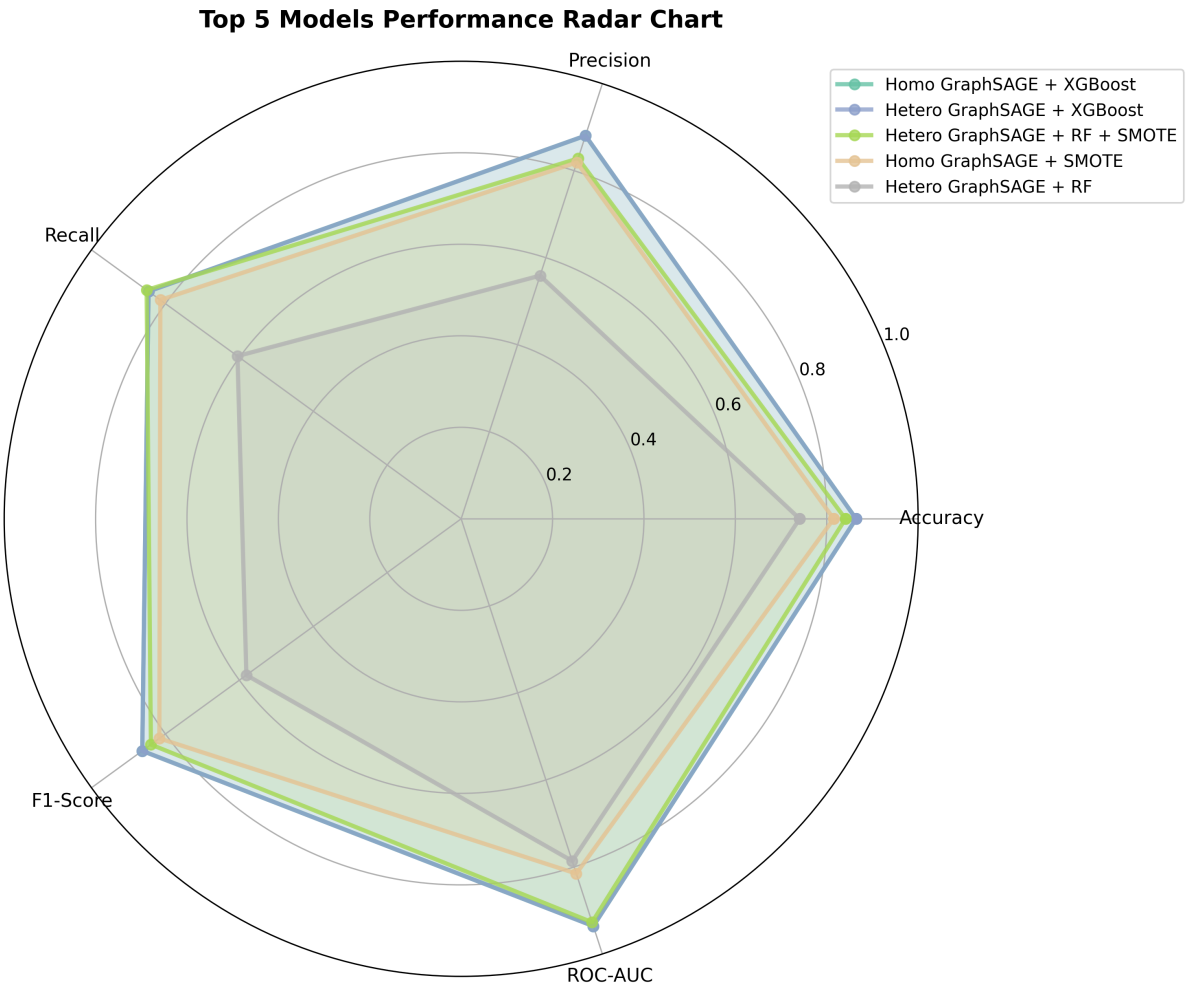

Figure 5: Top 5 models based on their Performance

produced strong performance (accuracy 0.8085, AUC-ROC 0.8663), consistent with the wider view that gradient-boosted trees are highly competitive on structured churn features. Feature attribution from SHAP indicated that tenure, contract length, fibre-optic internet service, monthly charges, and total charges dominate churn risk, which is operationally plausible as they reflect customer lifecycle position, lock-in, and price–value tension.

5.2 Value of Relational Learning (GraphSAGE) and Why It Alone Was Not Enough

The heterogeneous GraphSAGE model leveraged a multi-relation graph (16 node types and 15 edge types) to encode relational context around customers. Its performance (accuracy 0.7972, AUC-ROC 0.8515) showed that relational structure contains signal, but the model did not surpass XGBoost on discrimination. This outcome aligns with prior expectations: GraphSAGE is a strong inductive encoder for neighbourhood aggregation, but the downstream decision boundary can be harder to calibrate under imbalance and proxy-graph noise. When SMOTE was applied, GraphSAGE achieved a high recall of 0.8555 (with accuracy 0.7218), demonstrating that imbalance interventions can "push" the model toward catching churners at the cost of more false positives with an explicit trade-off that must be governed by campaign capacity and cost.

5.3 Why the Hybrid (GraphSAGE + XGBoost) Worked Best, and What It Means Practically

The strongest evidence for the research question came from the two-stage hybrid: GraphSAGE embeddings were fused with tabular inputs and learned by XGBoost, reaching accuracy 0.8641, F1 0.8627, and AUC-ROC 0.9360. This improvement over standalone XGBoost (AUC-ROC 0.8663) indicates complementary information: embeddings likely capture exposure/similarity effects not fully expressible through one-hot tabular columns. In contrast, GraphSAGE + Random Forest underperformed (accuracy 0.7408, AUC-ROC 0.7863), suggesting that not all downstream learners exploit embeddings equally well and that boosting's additive error-correction may be particularly suitable for embedding-augmented churn classification. For telecom application, this hybrid supports a clearer retention workflow: XGBoost provides feature-level explanations through SHAP, while the graph component can support relationship- or neighbourhood-based narratives, evidenced by the generation of relationship explanations for selected nodes. However, the business interpretation should ultimately be cost-sensitive; churn targeting is best framed as profit-maximising intervention rather than pure classification, so future evaluation should extend toward profit-based measures commonly advocated in churn analytics.

5.4 Limitations, Recommendations, and Concluding Statement

The main limitations follow directly from proxy graph construction and operational realism. If edges represent similarity rather than observed social ties, homophily and influence can be confounded, weakening causal interpretations of "peer effects" even when prediction improves. Graph explainability also remained partially constrained by implementation issues (e.g., GPU-related constraints on some interpretability steps), reinforcing

that explainability in hybrid systems is still a practical engineering challenge, not just a conceptual one.

References

AbdelAziz, N. M., Bekheet, M., Salah, A., El-Saber, N., & AbdelMoneim, W. T. (2025). A comprehensive evaluation of machine learning and deep learning models for churn prediction. *Information*, *16*(7), 537.

Barsotti, A. (2024). Prediction and causal analysis of churn in the telecommunication domain.

BlastChar. (n.d.). Telco customer churn [Accessed: 2025-12-10]. https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Chang, V., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., & Benson, V. (2024). Prediction of customer churn behavior in the telecommunication industry using machine learning models. *Algorithms*, *17*(6), 231.

Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). Customer churn prediction in telecommunication industry using deep learning. *Information Sciences Letters*, *11*(1), 24.

Grover, A., & Leskovec, J. (2016). Node2vec: Scalable feature learning for networks. https://doi.org/10.48550/arXiv.1607.00653

Hamilton, W. L., Ying, R., & Leskovec, J. (2018). Inductive representation learning on large graphs. https://doi.org/10.48550/arXiv.1706.02216

Hao, M. (2024). Telecom customer churn prediction based on bigru-attention-xgboost model. *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, 1050–1054. https://ieeexplore.ieee.org/abstract/document/10581856/

Hermawan, A., Saputra, A., Rafi, M. D., Basmallah, S., Putra, Y. T. S., & Nabila, W. (2025). Implementing xgboost model for predicting customer churn in e-commerce platforms. *Repeater: Publikasi Teknik Informatika Dan Jaringan*, *3*(2), 17–31.

Imani, M., Hashmi, D., & Verma, S. (2025). Evaluating the performance of random forest and xgboost with gaussian noise upsampling technique for customer churn prediction.

JM, A. K., & Dhanushkumar, V. (2025). Real-time customer churn prediction in telecommunication. *2025 International Conference on Data Science and Business Systems (ICDSBS)*, 1–7. https://ieeexplore.ieee.org/abstract/document/11031719/

Liu, R., Ali, S., Bilal, S. F., Sakhawat, Z., Imran, A., Almuhaimeed, A., Alzahrani, A., & Sun, G. (2022). An intelligent hybrid scheme for customer churn prediction integrating clustering and classification algorithms. *Applied Sciences*, *12*(18), 9355.

Miao, M., Miao, T., & Long, Z. (2024). User churn prediction hierarchical model based on graph attention convolutional neural networks. *China Communications*, *21*(7), 169–185.

Museba, T. (2024). A machine learning approach for customer churn prediction. *2024 4th International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 400–406. https://ieeexplore.ieee.org/abstract/document/10851064/

Noviandy, T. R., Idroes, G. M., Hardi, I., Afjal, M., & Ray, S. (2024). A model-agnostic interpretability approach to predicting customer churn in the telecommunications industry. *Infolitika Journal of Data Science*, *2*(1), 34–44.

Odusami, M., Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., & Sharma, M. M. (2021). A hybrid machine learning model for predicting customer churn in the telecommunication industry. In A. Abraham, H. Sasaki, R. Rios, N. Gandhi, U. Singh & K. Ma (Eds.), *Innovations in bio-inspired computing and applications* (pp. 458–468, Vol. 1372). Springer International Publishing. https://doi.org/10.1007/978-3-030-73603-3_43

Ouf, S., Mahmoud, K. T., & Abdel-Fattah, M. A. (2024). A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry. *Journal of Big Data*, *11*(1), 70. https://doi.org/10.1186/s40537-024-00922-9

Parmar, P., & Serasiya, S. (2021). Telecom churn prediction model using xgboost classifier and logistic regression algorithm. *International Research Journal of Engineering and Technology (IRJET)*, *8*, 1100–1105.

Peng, K., & Peng, Y. (2022). Research on telecom customer churn prediction based on ga-xgboost and shap. *Journal of Computer and Communications*, *10*(11), 107–120.

Peng, K., Peng, Y., & Li, W. (2023). Research on customer churn prediction and model interpretability analysis. *PLOS ONE*, *18*(12), e0289724.

Poudel, S. S., Pokharel, S., & Timilsina, M. (2024). Explaining customer churn prediction in telecom industry using tabular machine learning models. *Machine Learning with Applications*, *17*, 100567.

Senthilselvi, A., Kanishk, V., Vineesh, K., & Raj, A. P. (2024). A novel approach to customer churn prediction in telecom. *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, 1–7. https://ieeexplore.ieee.org/abstract/document/10602345/

Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., & Hjelm, R. D. (2018). Deep graph infomax. https://doi.org/10.48550/arXiv.1809.10341