

Project Dissertation Report on
Predicting Customer Churn Using Machine
Learning Techniques: A Case Study on Telecom
Customers

Submitted By :

Aashish Kumar

Under the Guidance of :

Dr. Yogesh Sharma

Assistant Professor



DELHI SCHOOL OF MANAGEMENT

Delhi Technological University

Bawana Road Delhi 110042

CERTIFICATE

This is to certify that the project titled “Predicting Customer Churn using Machine Learning techniques: A case study on Telecom Customers” is a bona fide work carried out by Aashish Kumar under my guidance in partial fulfilment of the requirements for the award of the degree of MBA at Delhi Technological University.

Date:

Place:

(Signature of Guide)

Name: Dr. Yogesh Sharma

Designation: Assistant Professor

DECLARATION

I hereby declare that the project report titled “Predicting Customer Churn using Machine Learning techniques: A case study on Telecom Customers” is my original work and has been carried out under the guidance of my faculty mentor.

Date:

Place:

(Signature)

Name: Aashish Kumar

Roll No: 24/BMBA/01

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my faculty guide for their valuable guidance, continuous support, and encouragement throughout the completion of this project.

I am also thankful to the Delhi School of Management, Delhi Technological University, for providing the necessary resources and academic environment to carry out this research work.

Lastly, I would like to thank my family and friends for their constant support and motivation.

Aashish Kumar

24/BMBA/01

EXECUTIVE SUMMARY

The telecommunications industry faces one of the most persistent and financially consequential challenges in customer management: the prediction and prevention of customer churn. Churn, defined as the voluntary discontinuation of a subscription service, directly impacts revenue stability, inflates customer acquisition costs, and undermines long-run profitability. Industry research consistently estimates that acquiring a new customer costs five to seven times more than retaining an existing one, making every prevented defection a measurable financial contribution to the firm's bottom line. As network infrastructure becomes increasingly commoditized and service offerings difficult to differentiate on technical grounds alone, the strategic focus of leading telecom operators has shifted decisively from acquisition to retention, placing churn prediction at the center of competitive customer management strategy.

This dissertation investigates the application of machine learning classification techniques to telecom customer churn prediction using the IBM Telco Customer Churn dataset sourced from Kaggle, comprising 7,043 customer records and 21 features spanning demographic attributes, account characteristics, service configuration variables, and financial metrics. Three classifiers — Logistic Regression, Decision Tree, and XGBoost — are evaluated under both imbalanced baseline conditions and three SMOTE-based resampling configurations: Standard SMOTE, Borderline-SMOTE, and SVMSMOTE. Systematic hyperparameter tuning using RandomizedSearchCV with five-fold stratified cross-validation is applied to optimize each configuration for minority-class detection performance. Model performance is assessed using five metrics — Accuracy, Precision, Recall, F1-score, and ROC-AUC — with Recall and F1-score designated as primary evaluation criteria, consistent with the operational objective of maximizing the identification of genuine at-risk customers.

Exploratory data analysis identifies contract type, customer tenure, internet service type, and value-added service subscriptions as the primary structural drivers of churn propensity.

Month-to-month customers churn at 42.7 percent compared to only 2.8 percent for two-year contract customers — a ratio exceeding fifteen-to-one that underscores the enormous protective effect of long-term contractual commitment. Newly acquired customers in their first twelve months of service exhibit the highest attrition risk, while subscribers to value-added services such as online security, technical support, online backup, and device protection consistently show lower churn rates across all customer segments, consistent with the switching cost hypothesis that each additional service increases the complexity and cost of switching to a competitor. Senior citizen customers exhibit disproportionately elevated churn rates of approximately 41.7 percent compared to 23.6 percent for non-senior customers, identifying them as a structurally distinct high-risk segment warranting dedicated retention attention.

Baseline model results confirm the severity of the class imbalance problem. All three classifiers trained on the raw imbalanced dataset achieved accuracy values near 0.80 — figures that are almost entirely explained by the correct classification of the dominant non-churn majority rather than genuine predictive capability. Recall values for the minority churn class ranged from only 0.497 to 0.550, meaning that between 45 and 50 percent of actual churners were misclassified as non-churners. A deployed model performing at this level would fail to flag nearly half of all customers who subsequently leave, rendering any retention program built on its outputs severely incomplete and operationally ineffective.

The application of SMOTE-based resampling produced substantial and consistent improvements in minority-class performance across all three classifiers, with Recall gains ranging from 18 to 29 percentage points depending on the classifier-resampling combination. Among all twelve model-resampling configurations evaluated, XGBoost combined with SVM SMOTE achieved the strongest overall performance — an F1-score of 0.636, Recall of 0.758, and ROC-AUC of 0.845. Compared to the baseline XGBoost model, the SVM SMOTE-balanced configuration reduced missed churners by 52 percent, from 235 to 113 False Negatives, while increasing correctly identified churners from 232 to 354 True Positives. The increase in False Positives from 115 to 293 represents an operationally acceptable trade-off, as the cost of incorrectly targeting a non-churner with a retention offer is substantially lower than the cost of failing to intervene with a genuine at-risk customer.

The study concludes with five principal findings: class imbalance is a genuine and consequential methodological problem that accuracy-focused evaluation cannot adequately capture; SMOTE-based resampling — SVM SMOTE in particular — is a mandatory rather than optional component of any production churn prediction pipeline; XGBoost with

SVMSMOTE is the recommended configuration for deployment, delivering the strongest F1-score, most stable ROC-AUC, and greatest reduction in missed churners; Logistic Regression with SVMSMOTE represents a competitive interpretability-friendly alternative for organizations with regulatory or stakeholder transparency requirements; and contract type combined with early tenure represents the highest-value target for proactive retention intervention. Together, these findings provide both a validated analytical framework and actionable strategic guidance for telecom operators seeking to implement effective, data-driven customer retention programs.

TABLE OF CONTENTS

S.NO	Topic	Page No.
1.	Certificate	ii
2.	Declaration	iii
3.	Acknowledgement	iv
4.	Executive Summary	v
5.	Table of Contents	viii
6.	Introduction	1
7.	Literature Review	8
8.	Research Methodology	16
9.	Data Analysis & Results	27
10.	Conclusion	46
11.	References	52
13.	Annexure	55

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The telecommunications industry occupies a foundational position in the global digital economy. As connectivity has transitioned from a convenience to an essential infrastructure, telecom operators have grown into some of the largest and most complex service businesses in the world. In mature markets, where network coverage is near-universal and the technical differentiation between competing providers has narrowed significantly, the primary battleground of competitive strategy has shifted decisively from acquisition to retention. The central managerial challenge facing telecom firms today is not how to attract new customers, but how to keep the ones they already have.

This shift in strategic emphasis has given rise to a discipline of growing analytical sophistication: the prediction and prevention of customer churn. Churn, defined as the voluntary or involuntary discontinuation of a customer's subscription to a telecom service, represents one of the most direct and measurable threats to revenue stability in the industry. A customer who leaves takes with them not only the revenue of the current billing period but the entire stream of future payments that would have accumulated over the remaining duration of a long-term relationship. The financial asymmetry between retention and acquisition further compounds this loss: industry estimates consistently place the cost of acquiring a new customer at five to seven times the cost of retaining an existing one, making churn prevention not merely a tactical concern but a strategic priority with material implications for profitability, valuation, and long-run competitive positioning.

The intellectual and analytical response to this challenge has evolved considerably over the past three decades. Early approaches to churn management relied primarily on demographic segmentation and simple rule-based systems, identifying at-risk customers based on observable characteristics such as contract type, tenure, or billing patterns. While these methods provided a starting point, they suffered from a fundamental limitation: they captured correlation without capturing complexity. Customer churn is not a simple function of any single variable or even a small set of variables. It emerges from the interaction of multiple

behavioural, financial, and service-quality signals, many of which influence each other in ways that simple linear models cannot represent.

The emergence of machine learning as a mainstream analytical toolkit over the past two decades has transformed what is possible in churn prediction. Algorithms such as Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting methods including XGBoost have demonstrated the capacity to learn complex, non-linear patterns from large customer datasets, producing predictions that are substantially more accurate than what traditional statistical methods could achieve. Their adoption in telecom analytics has been rapid and broad, driven by the consistent empirical finding that churn propensity is a non-linear, multi-dimensional phenomenon that rewards analytical sophistication.

However, the practical deployment of machine learning models in churn prediction has encountered a persistent and structurally important obstacle that the academic literature has been slow to fully address: the problem of class imbalance. In virtually every real-world churn dataset, the proportion of customers who churn in any given period is substantially smaller than the proportion who do not. In the IBM Telco Customer Churn dataset that forms the empirical foundation of this study, approximately 26 percent of customers are churners and 74 percent are non-churners. This imbalance, while modest by the standards of some domains, is sufficient to severely distort the behavior of standard machine learning algorithms. Models trained on imbalanced data learn a systematic bias toward the majority class, producing predictions that appear accurate in aggregate because they correctly classify the dominant non-churn population, but perform poorly on the minority churn class that is actually of greatest interest. A model that predicts "no churn" for every customer would achieve accuracy close to 74 percent on this dataset while being entirely useless for its intended purpose.

Addressing class imbalance has therefore become a central methodological challenge in the churn prediction literature. A family of techniques collectively known as Synthetic Minority Oversampling Technique, or SMOTE, has emerged as the dominant approach, generating synthetic examples of the minority class to balance the training distribution and enable models to learn minority-class characteristics effectively. The original SMOTE algorithm has been extended and refined through several variants, including Borderline-SMOTE, which focuses synthetic generation on the most difficult boundary cases, and SVMSMOTE, which uses a Support Vector Machine to identify the optimal regions for synthetic sample placement.

The comparative effectiveness of these variants across different model architectures has not been systematically studied in the telecom churn context, representing a meaningful gap that this study addresses.

The IBM Telco Customer Churn dataset, sourced from Kaggle and originally published by IBM, provides a rich empirical environment for this investigation. The dataset encompasses 7,043 customer records and 21 features spanning demographic characteristics, contractual arrangements, service configurations, and financial attributes. Its breadth of variables enables the kind of multi-dimensional churn analysis that single-feature approaches cannot support, and its widespread use in the research community provides a benchmark against which the findings of this study can be contextualized.

This study was conducted as part of the Machine Learning course (MB 302) in the MBA Business Analytics program at Delhi School of Management, Delhi Technological University. It investigates the comparative performance of three machine learning classifiers — Logistic Regression, Decision Tree, and XGBoost — under both imbalanced and SMOTE-balanced conditions, with the objective of identifying the model configuration that best serves the practical needs of telecom customer retention management.

1.2 Problem Statement

The prediction of customer churn in the telecommunications industry is characterized by four interconnected problems, each of which limits either the accuracy, the reliability, or the practical utility of existing approaches. This study is organized around the simultaneous resolution of all four.

The Class Imbalance Problem. Standard machine learning classifiers are designed under the implicit assumption that the classes being predicted are approximately equally represented in the training data. When this assumption is violated, as it consistently is in churn datasets where churners represent a minority of the customer base, the classifier's optimization process is distorted. By minimizing the overall error rate, the algorithm learns to favor the majority class, achieving high aggregate accuracy at the cost of poor sensitivity to the minority class. In churn prediction, this trade-off is strategically unacceptable: the minority class the churners is precisely the population about which accurate predictions are needed. A model that

achieves 74 percent accuracy by predicting no churn for every customer provides zero actionable value, yet standard evaluation metrics such as accuracy would rate it favorably.

Correcting this imbalance requires both appropriate resampling techniques applied during training and appropriate evaluation metrics that capture minority-class performance directly.

The Algorithm Selection Problem. The landscape of machine learning algorithms available for classification tasks is broad, and different algorithms make different assumptions about the structure of the data-generating process. Logistic Regression assumes linearity in the log-odds of the outcome; Decision Trees partition the feature space through recursive binary splits; XGBoost constructs an additive ensemble of weak learners through sequential gradient-based optimization. Each architecture has different strengths and weaknesses in the context of churn prediction, and the choice among them is not straightforward. Empirical comparison under controlled conditions is necessary to establish which algorithm, with which resampling configuration, produces the most reliable and actionable churn predictions on telecom data.

The Evaluation Validity Problem. A significant portion of the churn prediction literature reports performance metrics that do not accurately reflect the real-world utility of the models described. The most common source of this distortion is the use of accuracy as the primary evaluation criterion on imbalanced datasets, which, as noted above, rewards the trivial strategy of predicting the majority class. A rigorous evaluation framework for churn prediction must prioritize metrics that capture minority-class performance: recall, which measures the proportion of actual churners correctly identified; F1-score, which balances recall and precision; and ROC-AUC, which assesses the model's discriminative ability across all possible classification thresholds. Only when evaluated on these metrics can the true relative performance of competing model configurations be accurately assessed.

The Deployment Gap Problem. Even a model that is accurate, properly evaluated, and well-calibrated for minority-class detection generates business value only if its outputs can be translated into specific, actionable retention interventions. A churn probability score attached to each customer record is a starting point, not an endpoint. The practical value of churn prediction is realized through targeted outreach: identifying the customers most likely to leave and deploying retention resources, whether personalized offers, proactive service quality interventions, or loyalty program benefits, in proportion to the predicted risk and the estimated value of the relationship. Bridging the gap between statistical prediction and

operational marketing strategy requires both a model with sufficient recall to surface most at-risk customers and a framework for translating predicted probabilities into prioritized action.

This study addresses all four problems within a single integrated empirical framework, using the IBM Telco Customer Churn dataset, a systematic comparison of SMOTE variants, a multi-metric evaluation protocol, and a set of business recommendations grounded directly in the empirical findings.

1.3 Objectives of the Study

The study is guided by the following research objectives, progressing from data preparation through modeling to business application:

Objective 1: To conduct a thorough exploratory analysis of the IBM Telco Customer Churn dataset, characterizing the distributional properties of all features, identifying patterns associated with churn behavior, and documenting the nature and extent of class imbalance in the target variable.

Objective 2: To implement a rigorous data preprocessing pipeline encompassing missing value treatment, feature encoding, feature scaling, and stratified train-test splitting, ensuring that all subsequent model training and evaluation is conducted on properly prepared data under consistent conditions.

Objective 3: To train and evaluate three machine learning classifiers — Logistic Regression, Decision Tree, and XGBoost — on the imbalanced dataset as baseline configurations, establishing reference performance levels against which the effects of resampling can be measured.

Objective 4: To apply three SMOTE-based oversampling techniques — Standard SMOTE, Borderline-SMOTE, and SVMSMOTE — to the training data and retrain all three classifiers under each resampling configuration, producing a comprehensive matrix of model-resampling combinations for comparative analysis.

Objective 5: To evaluate all model configurations using a multi-metric framework encompassing Accuracy, Precision, Recall, F1-score, and ROC-AUC, with particular emphasis on Recall and F1-score as the primary criteria appropriate to the churn prediction objective.

Objective 6: To conduct hyperparameter tuning using RandomizedSearchCV across all three classifiers, optimizing each model's configuration for maximum minority-class performance while maintaining generalization to held-out test data.

Objective 7: To identify the optimal model-resampling configuration through systematic empirical comparison and to provide evidence-based business recommendations for telecom operators seeking to implement the resulting churn prediction system in a production CRM environment.

1.4 Scope of the Study

The boundaries of this study are defined along four dimensions — data, features, models, and context — and these boundaries are maintained consistently throughout the empirical work.

Data Scope. The analysis draws exclusively on the IBM Telco Customer Churn dataset as sourced from Kaggle. The dataset represents a cross-sectional snapshot of a single telecom operator's customer base at a specific point in time, encompassing 7,043 unique customer records. It does not represent a longitudinal panel, which means that the temporal evolution of individual customer behavior cannot be tracked within this dataset. The findings are most directly applicable to telecom operators with similar customer structures and service configurations, and should not be assumed to generalize without validation to substantially different industries, geographies, or customer segments.

Feature Scope. The predictive features employed are restricted to those present in the IBM Telco dataset: seven demographic variables (gender, senior citizen status, partner, dependents), five account variables (tenure, contract type, paperless billing, payment method), eight service variables (phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming TV and movies), and two financial variables (monthly charges, total charges). The target variable is the binary churn indicator. No external data sources, behavioral signals beyond those captured in the dataset, or proprietary features are incorporated. This restriction ensures full reproducibility of the study's methodology and findings.

Model Scope. The modeling work covers three classification algorithms selected to span a deliberate range of complexity, interpretability, and expected performance: one parametric linear model (Logistic Regression), one non-parametric rule-based model (Decision Tree), and one high-performance gradient boosting ensemble (XGBoost). Deep learning

architectures, including recurrent networks and attention-based models suitable for sequential customer interaction data, are outside the scope of this study. While such methods have shown promise in churn prediction research, their data requirements, computational overhead, and relative opacity make them less suitable for the interpretability-conscious, business-facing application context targeted here.

Resampling Scope. The study evaluates three oversampling methods from the SMOTE family: Standard SMOTE, Borderline-SMOTE, and SVMSMOTE. Undersampling techniques, cost-sensitive learning approaches, and hybrid methods combining oversampling with undersampling are not evaluated in this study, representing directions for future research. The resampling procedures are applied exclusively to the training data, with the test set maintained in its original imbalanced form to ensure that performance estimates reflect real-world deployment conditions.

Evaluation Scope. Model performance is assessed using five standard metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC. While more specialized metrics such as the Matthews Correlation Coefficient or cost-weighted performance measures could provide additional perspectives, the chosen metrics represent the standard evaluation framework used across the churn prediction literature and provide sufficient information for the comparative conclusions of this study.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction to the Literature

The prediction of customer churn has attracted sustained academic and practitioner attention for over two decades, driven by the strategic and financial consequences of customer attrition in subscription-based industries. The telecommunications sector, characterized by high customer acquisition costs, intense competitive pressure, and commoditized service offerings, has been a particularly productive domain for churn prediction research. The literature in this area has evolved through several distinct phases: early rule-based and statistical approaches, the adoption of classical machine learning classifiers, the integration of ensemble methods, and most recently, the systematic treatment of class imbalance as a first-order methodological concern. This chapter traces this evolution, situates the present study within the current state of the literature, and identifies the specific gaps that the empirical work addresses.

2.2 Early Approaches to Churn Prediction

The earliest formal treatments of customer churn in the marketing and management literature approached the problem through the lens of customer relationship theory. Reichheld and Sasser (1990) provided one of the foundational arguments for the economic importance of retention, demonstrating empirically that even modest improvements in customer retention rates could produce substantial increases in firm profitability, owing to the cumulative revenue streams generated by long-term customers and the disproportionate cost of replacing them. This work established the strategic context within which churn prediction would later be developed as an analytical discipline.

The initial analytical approaches to churn prediction relied primarily on classical statistical methods. Logistic Regression emerged early as a natural candidate for binary churn classification, owing to its theoretical transparency, computational tractability, and well-understood probabilistic interpretation. Early applications of logistic regression to telecom churn data, reviewed comprehensively by Hadden et al. (2006), demonstrated that while the

method produced interpretable coefficient estimates that connected specific customer characteristics to churn probability, its predictive accuracy was constrained by its linearity assumption. Customer churn behavior, even in its earliest studied forms, exhibited interactions and non-linearities that logistic regression could not fully represent.

Survival analysis methods, borrowed from biostatistics and reliability engineering, offered an alternative framework that was theoretically well-suited to the churn problem. Cox proportional hazards models and accelerated failure time models treated the time until churn as the outcome of interest, allowing the incorporation of customer tenure and the modeling of time-varying covariates. Bock et al. (2005) demonstrated that survival analysis approaches could outperform logistic regression on datasets where the temporal dimension of churn was an important source of predictive signal. However, their complexity of implementation and the difficulty of communicating their outputs to non-technical business stakeholders limited their adoption in commercial settings.

Decision Trees represented another early entrant into the churn prediction toolkit. Their hierarchical partitioning of the feature space produced human-readable decision rules that could be directly inspected and operationalized by marketing teams, offering a degree of transparency that more complex models lacked. Mozer et al. (2000) applied decision tree methods to telecom churn prediction and found that the resulting rule sets captured meaningful behavioral patterns, though their accuracy was generally inferior to ensemble methods that would be developed subsequently. The tendency of single decision trees to overfit the training data, producing highly accurate but poorly generalizing models, was identified early as their primary limitation.

2.3 Machine Learning Methods in Churn Prediction

The systematic application of machine learning to churn prediction accelerated significantly in the 2000s, driven by the increasing availability of rich customer databases, the development of practitioner-friendly software libraries, and the consistent empirical finding that non-linear methods outperformed classical statistical approaches on real-world churn datasets.

Random Forest, introduced by Breiman (2001), represented a major advance in the treatment of the overfitting problem that plagued single decision trees. By training an ensemble of trees on bootstrapped samples of the training data and averaging their predictions, Random Forest

achieved a bias-variance balance that proved difficult to replicate with simpler methods. Its robustness to irrelevant features and its ability to naturally handle interactions between predictors made it particularly well-suited to churn datasets, which typically contain a mixture of informative and redundant variables across demographic, behavioral, and financial dimensions. Verbeke et al. (2012) conducted one of the most comprehensive early comparative studies of machine learning methods for churn prediction, evaluating ten classifiers across multiple telecom datasets and finding that ensemble methods including Random Forest consistently outperformed single classifiers across the majority of evaluation criteria.

Gradient Boosting methods, originating with the theoretical framework established by Friedman (2001), approached the ensemble problem from a different direction. Rather than reducing variance through parallel aggregation of independently trained trees, gradient boosting reduced bias through sequential correction, with each new tree in the ensemble fitted to the residuals of the current model. The resulting models, while more computationally intensive and more sensitive to hyperparameter choices than Random Forest, achieved state-of-the-art predictive accuracy on structured tabular data across a wide range of classification tasks. Their application to telecom churn prediction demonstrated consistent accuracy advantages over both single classifiers and earlier ensemble methods.

XGBoost, developed by Chen and Guestrin (2016), represented a highly optimized implementation of the gradient boosting framework that introduced several important refinements: regularization terms in the objective function to control overfitting, a block structure for parallel computation, and an efficient algorithm for handling sparse data. These enhancements produced a method that was both faster and more accurate than earlier gradient boosting implementations on most benchmark tasks. The rapid adoption of XGBoost across competitive machine learning benchmarks, including multiple Kaggle competition victories on tabular data problems, established it as a leading candidate for churn prediction applications. Manzoor et al. (2024) conducted a systematic review of over 200 churn prediction papers published between 2015 and 2023 and found that XGBoost-based approaches appeared among the top-performing configurations in a majority of comparative studies across multiple industries, with the telecom sector representing the largest single domain of application.

Support Vector Machines (SVM) were also explored extensively in the churn prediction literature during the 2000s and early 2010s. Vafeiadis et al. (2015) conducted a comparative evaluation of multiple machine learning techniques for customer churn prediction, including

SVM, Naive Bayes, Artificial Neural Networks, and Decision Trees, and found that SVM achieved competitive performance on several benchmark datasets. However, SVM's computational cost on large datasets, its sensitivity to feature scaling, and the difficulty of interpreting its predictions in business terms limited its adoption relative to tree-based methods in practical deployments.

Neural network and deep learning approaches have been explored more recently in the churn prediction literature, motivated by their demonstrated success in image recognition, natural language processing, and other high-dimensional pattern recognition tasks. Ma and Xia (2019) applied deep neural networks to customer churn prediction and reported accuracy improvements over traditional machine learning methods on specific datasets. However, the data requirements, computational cost, and near-total opacity of deep learning architectures have constrained their adoption in churn prediction applications, where datasets are typically moderate in size and interpretability is valued for operational deployment.

2.4 The Class Imbalance Problem in Churn Prediction

The recognition of class imbalance as a first-order methodological challenge in churn prediction emerged gradually through the early 2000s and has since become one of the most actively researched topics in the domain. The practical significance of the problem is straightforward: in virtually every real-world churn dataset, churners represent a minority of the total customer base, typically ranging from 5 to 30 percent depending on the industry, the measurement period, and the operational definition of churn. Standard machine learning algorithms, optimized to minimize overall classification error, respond to this imbalance by systematically under-predicting the minority class, producing models that are highly accurate in aggregate but nearly useless for identifying the customers most in need of retention intervention.

Chawla et al. (2002) introduced the Synthetic Minority Oversampling Technique as a principled response to this problem. Rather than simply replicating existing minority-class examples, which would not provide the classifier with new information, SMOTE generates synthetic examples by interpolating between pairs of existing minority-class instances in feature space. For each minority-class example, a set of its nearest neighbors is identified, and new synthetic examples are created by randomly selecting points along the line segments connecting the original example to its neighbors. The resulting augmented training set provides the classifier with a richer and more varied representation of the minority class,

enabling it to learn decision boundaries that are more sensitive to minority-class characteristics.

The original SMOTE algorithm has been refined and extended through a series of variants that address specific limitations of the basic approach. Han et al. (2005) introduced Borderline-SMOTE, which concentrates synthetic sample generation on the minority-class examples that lie closest to the decision boundary, the instances that are most difficult to classify correctly and therefore most informative for boundary learning. The intuition is that synthetic examples generated far from the decision boundary provide limited additional information to the classifier, while examples near the boundary help define the classification surface more precisely.

SVMSMOTE, proposed by Nguyen et al. (2009), takes a different approach to identifying the optimal regions for synthetic sample generation, using a Support Vector Machine to characterize the decision boundary and focusing synthetic generation on the minority-class instances that lie within the SVM's margin region. This approach produces synthetic examples that are more informative for classifying difficult boundary cases and has been shown to outperform both Standard SMOTE and Borderline-SMOTE on several benchmark datasets with complex decision boundaries.

Dhanawade (2023) applied SMOTE in combination with multiple machine learning classifiers to telecom churn prediction and demonstrated that synthetic oversampling consistently improved recall for the churn class across all classifiers tested, with the magnitude of improvement varying with both the classifier architecture and the specific SMOTE variant employed. Suguna and Premalatha (2025) extended this analysis by combining SMOTE with ensemble methods and demonstrating that the interaction between oversampling strategy and classifier architecture produced non-trivial effects on F1-score and ROC-AUC, with certain combinations outperforming others in ways that could not be predicted from the performance of either component in isolation. Sikri and Bhat (2024) introduced a ratio-based data balancing technique and integrated it with XGBoost, demonstrating competitive performance against standard SMOTE variants and highlighting the sensitivity of churn prediction results to the specific imbalance handling strategy employed.

2.5 Comparative Studies and Ensemble Methods

The comparative evaluation of machine learning methods for churn prediction has been a consistent focus of the literature, motivated by the practical need for evidence-based model selection guidance. Urrahman and Jawad (2024) conducted a comparative analysis of churn prediction techniques across multiple telecom datasets and found that ensemble models, particularly XGBoost and Random Forest, consistently outperformed traditional classifiers such as Logistic Regression and Support Vector Machines, especially when coupled with oversampling techniques. This finding was consistent with the broader pattern identified by Manzoor et al. (2024) in their systematic review, which found ensemble methods appearing among the top performers in a majority of comparative studies.

However, comparative studies have also identified important nuances in the relationship between model complexity and predictive performance. Wagh (2024) demonstrated that XGBoost's advantage over simpler methods was sensitive to the quality of hyperparameter tuning, with poorly configured XGBoost models sometimes underperforming well-tuned logistic regression baselines. This finding underscores the importance of systematic hyperparameter optimization as a component of rigorous comparative evaluation, rather than comparing models in their default configurations. Lu and Wang (2021) further highlighted the role of feature selection in determining comparative model performance, demonstrating that the accuracy advantage of complex ensemble methods over simpler alternatives could be substantially reduced when both were provided with an optimally selected feature subset.

The specific combination of XGBoost with SVM SMOTE has received relatively limited systematic evaluation in the published literature despite its strong theoretical motivation. The Journal of Engineering Research and Reports (2025) identified this combination as a promising direction for telecom churn prediction but did not provide a comprehensive empirical evaluation across multiple SMOTE variants and classifier architectures. The present study fills this gap by providing exactly such an evaluation.

2.6 Evaluation Frameworks for Imbalanced Classification

The selection of appropriate evaluation metrics for imbalanced classification tasks has been a subject of considerable methodological debate in the machine learning literature. The fundamental inadequacy of accuracy as a primary evaluation criterion for imbalanced datasets is well established: a classifier that predicts the majority class for every instance achieves

accuracy equal to the majority class proportion while providing zero information about minority-class behavior. Despite this well-known limitation, accuracy continues to appear as a primary metric in a substantial portion of the churn prediction literature, creating misleadingly optimistic impressions of model performance.

The F1-score, defined as the harmonic mean of precision and recall, has emerged as the preferred summary metric for imbalanced classification in the churn prediction context, because it simultaneously captures both the proportion of actual churners identified (recall) and the proportion of predicted churners who are genuine (precision). For churn prediction applications where the cost of missing an actual churner typically exceeds the cost of incorrectly targeting a non-churner with retention resources, recall is particularly important. Yan (2025) applied machine learning-based prediction to customer churn and emphasized that recall-focused evaluation was essential for assessing the practical utility of churn models in business settings, arguing that the business cost of false negatives in churn prediction systematically exceeds the cost of false positives.

The ROC-AUC metric, which measures the area under the Receiver Operating Characteristic curve, provides a threshold-independent assessment of the classifier's discriminative ability across all possible operating points. Unlike F1-score, which is computed at a single classification threshold, ROC-AUC captures the full range of precision-recall trade-offs available to a deployed system. Its use in the churn prediction literature has been advocated by multiple authors as a complement to threshold-dependent metrics, providing a more complete picture of model performance. The present study employs all five metrics — Accuracy, Precision, Recall, F1-score, and ROC-AUC — in a multi-criterion evaluation framework, with particular emphasis on Recall and F1-score as primary criteria.

2.7 Practical Deployment and Business Implications

A dimension of the churn prediction literature that has received less systematic attention than model development and evaluation is the translation of model outputs into actionable business strategy. The gap between a churn probability score and a concrete retention intervention is not trivial: it requires decisions about which customers to target, which interventions to offer, how to prioritize limited retention resources, and how to measure the incremental effectiveness of the program. Several authors have addressed aspects of this translation problem.

Wagh (2024) framed the business case for churn prediction explicitly in terms of the cost differential between retention and acquisition, arguing that even a modest improvement in churn prediction accuracy, if translated into appropriately targeted retention campaigns, could generate returns that substantially exceeded the cost of model development and deployment. The practical implication of this framing is that the relevant evaluation criterion for a business-facing churn model is not merely statistical accuracy but the incremental revenue preserved through more precise identification of at-risk customers.

The integration of churn prediction models into Customer Relationship Management systems has been identified by multiple authors as a critical enabler of operational deployment. Cloud-based deployment frameworks, real-time scoring pipelines, and automated intervention triggers have been discussed in the context of making churn prediction models actionable at scale. Urrahman and Jawad (2024) noted that the practical deployment of ensemble models such as XGBoost in CRM systems required investment in both technical infrastructure and organizational capability, with model interpretability tools such as SHAP values playing an important role in building the trust of business stakeholders who must act on the model's recommendations.

2.8 Research Gap

The preceding review identifies a convergence of underaddressed issues in the current churn prediction literature. First, while the individual components of the present study — SMOTE variants, XGBoost, multi-metric evaluation — have each been explored separately, a comprehensive comparative evaluation of all three SMOTE variants across all three classifiers on the IBM Telco dataset, with systematic hyperparameter tuning and a full multi-metric evaluation framework, has not been published. Second, the relative performance of SVMSMOTE versus Standard SMOTE and Borderline-SMOTE in the XGBoost context has not been systematically characterized. Third, the translation of model findings into specific, evidence-grounded business recommendations for telecom operators has been treated superficially in most comparative studies. This study addresses all three gaps within a single integrated empirical framework, contributing both methodological rigor and substantive practical guidance to the literature.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Philosophy and Design

This study is grounded in a quantitative, positivist research philosophy. Customer churn is treated as an objectively measurable binary outcome — a customer either discontinues their subscription within a defined observation period or they do not — and the study proceeds from the assumption that this outcome can be predicted with meaningful accuracy from observable customer characteristics using mathematical models. The research design is applied and experimental in character: three classification algorithms with different theoretical properties are applied to the same dataset under identical preprocessing and evaluation conditions, and their outputs are compared using standardized performance criteria. This design enables controlled inference about the relationship between algorithmic and resampling properties and predictive outcomes in the telecom churn prediction context.

The study follows a deductive logic, proceeding from theoretical propositions about model behavior — for example, that ensemble methods will outperform single classifiers on non-linear classification tasks, and that SMOTE-based resampling will improve minority-class recall at some cost to overall accuracy — to empirical tests of those propositions using real customer data. Where empirical results align with theoretical expectations, the study provides quantitative confirmation of established principles. Where they deviate, the study offers explanatory analysis grounded in the algorithmic properties of the models and the specific characteristics of the dataset.

The overall analytical workflow proceeds through five sequential stages: data acquisition and initial inspection, data preprocessing and feature engineering, baseline model training and evaluation, SMOTE-based resampling and comparative model training, and hyperparameter tuning with final performance assessment. Each stage is documented in sufficient detail to ensure full reproducibility of the methodology and findings.

3.2 Dataset Description

3.2.1 Source and Overview

The empirical analysis is based on the IBM Telco Customer Churn dataset, originally published by IBM and made publicly available through the Kaggle data science platform. The dataset documents the characteristics and subscription status of 7,043 customers of a fictional but realistically structured telecommunications company. Each record in the dataset represents a single unique customer and includes 20 predictor variables spanning four broad categories — demographic attributes, account information, service configuration, and financial metrics — along with a binary target variable indicating whether the customer churned during the observation period.

The dataset has been widely used as a benchmark in the churn prediction literature, providing a well-understood and extensively validated empirical environment. Its breadth of feature categories enables multi-dimensional churn analysis that captures the complex interplay of demographic, contractual, service-quality, and financial factors that collectively determine churn propensity. Its moderate size — 7,043 records — is representative of the customer databases maintained by small-to-medium telecom operators and ensures that the computational demands of model training and evaluation remain tractable.

3.2.2 Variable Descriptions

The dataset's 20 predictor variables are organized into four categories as follows:

Demographic Variables: Gender (binary: male/female), Senior Citizen status (binary: 0/1), Partner (binary: yes/no indicating whether the customer has a partner), and Dependents (binary: yes/no indicating whether the customer has dependents). These variables capture the household and life-stage characteristics that have been theoretically linked to service usage patterns and churn propensity in prior literature.

Account Variables: Tenure (numeric: number of months the customer has been with the company), Contract type (categorical: month-to-month, one-year, two-year), Paperless Billing (binary: yes/no), and Payment Method (categorical: electronic check, mailed check, bank transfer, credit card). These variables capture the nature and duration of the customer's contractual relationship with the firm, which prior research has identified as among the strongest structural predictors of churn.

Service Variables: Phone Service (binary: yes/no), Multiple Lines (categorical: yes/no/no phone service), Internet Service (categorical: DSL/fiber optic/no), Online Security

(categorical: yes/no/no internet service), Online Backup (categorical: yes/no/no internet service), Device Protection (categorical: yes/no/no internet service), Tech Support (categorical: yes/no/no internet service), Streaming TV (categorical: yes/no/no internet service), and Streaming Movies (categorical: yes/no/no internet service). These variables capture the breadth and configuration of the customer's service bundle, reflecting both usage intensity and switching cost considerations.

Financial Variables: Monthly Charges (numeric: the customer's current monthly billing amount in dollars) and Total Charges (numeric: the total amount billed to the customer across their tenure). These variables capture the financial dimension of the customer relationship and are expected to interact with contract type and tenure in complex ways.

Target Variable: Churn (binary: yes/no indicating whether the customer discontinued service). Customers coded as "yes" constitute the minority churn class of primary analytical interest.

3.2.3 Class Distribution

The target variable exhibits a class imbalance that is characteristic of real-world churn datasets. Of the 7,043 customer records, approximately 1,869 (26.5 percent) are churners and 5,174 (73.5 percent) are non-churners. This approximately 3:1 ratio of non-churners to churners is sufficient to materially bias standard classifiers toward the majority class and motivates the systematic application of SMOTE-based resampling techniques documented in Section 3.6.

3.3 Data Preprocessing

The raw dataset underwent a structured five-step preprocessing sequence before model training. All preprocessing steps were implemented in Python using the scikit-learn library, organized within a ColumnTransformer and Pipeline architecture that ensures consistent application of all transformations to both training and test data.

3.3.1 Missing Value Treatment

Inspection of the raw dataset revealed that the TotalCharges variable, which represents the cumulative billing amount across the customer's tenure, contained 11 records with blank string values rather than numeric entries. These blank values occurred exclusively among customers with a tenure of zero months, indicating that these customers were newly enrolled

at the time of the dataset snapshot and had not yet been billed. The TotalCharges variable was converted from string to numeric format using pandas' `to_numeric` function with the `errors='coerce'` parameter, which converted the blank strings to NaN values. These 11 records were subsequently removed from the dataset, as their missing TotalCharges values were not recoverable by imputation given that they represented genuinely unbilled customers rather than data entry errors. The removal of 11 records from a dataset of 7,043 represents a loss of 0.16 percent of observations, which is analytically negligible.

3.3.2 Non-Informative Variable Removal

The CustomerID variable, which serves as a unique administrative identifier for each customer record, was excluded from the analysis. As a randomly assigned identifier with no behavioral or financial content, CustomerID cannot carry predictive signal for churn and its inclusion would constitute noise in the feature space. The SeniorCitizen variable, originally coded as a binary integer (0/1), was retained in its numeric form.

3.3.3 Target Variable Encoding

The Churn target variable, originally coded as a binary string ("Yes"/"No"), was re-encoded as a binary integer (1/0) to conform to the requirements of scikit-learn classifiers and evaluation functions. Churn = "Yes" was mapped to 1 (the positive class representing churners) and Churn = "No" was mapped to 0 (the negative class representing retained customers).

3.3.4 Feature Encoding

Categorical predictor variables were transformed into numeric form using OneHotEncoder from scikit-learn's preprocessing module. OneHotEncoding creates a binary indicator column for each category level of each categorical variable, enabling classifiers that operate on numeric feature spaces to process categorical inputs without imposing an arbitrary ordinal structure on nominal variables. The `drop='first'` parameter was applied to avoid perfect multicollinearity by dropping one category level from each encoded variable, producing $k-1$ binary columns for a variable with k categories.

The categorical variables subjected to OneHotEncoding were: Gender, Partner, Dependents, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, and PaymentMethod.

3.3.5 Feature Scaling

Numeric predictor variables — Tenure, MonthlyCharges, and TotalCharges — were standardized using StandardScaler from scikit-learn's preprocessing module. StandardScaler transforms each numeric variable by subtracting its training-set mean and dividing by its training-set standard deviation, producing variables with zero mean and unit variance. This standardization is important for classifiers whose optimization processes are sensitive to the scale of input features, particularly Logistic Regression, whose gradient descent optimization converges more reliably on standardized inputs, and SVM SMOTE, whose nearest-neighbor calculations depend on Euclidean distances that are distorted by features measured on incomparable scales.

Critically, the scaling parameters — means and standard deviations — were computed exclusively from the training set and applied to both training and test sets. Computing scaling parameters from the full dataset before splitting would constitute a form of data leakage, incorporating test-set information into the preprocessing and producing overly optimistic performance estimates. This requirement is enforced by the scikit-learn Pipeline architecture, which ensures that all fitted transformations are learned exclusively from training data.

3.4 Train-Test Split

The preprocessed dataset was divided into training (75 percent) and testing (25 percent) subsets using scikit-learn's `train_test_split` function with the `stratify` parameter set to the target variable. Stratified splitting ensures that the class distribution of the target variable is preserved in both the training and test sets, maintaining the approximately 26/74 churn/non-churn ratio in each subset. This is important for ensuring that the test set provides an accurate representation of the real-world class distribution, and therefore that performance estimates computed on the test set reflect the model's expected behavior in deployment.

The resulting training set contains approximately 5,282 records and the test set contains approximately 1,761 records. A random seed was fixed at the point of splitting to ensure reproducibility of all subsequent analyses. The test set was held out entirely from all model training, hyperparameter tuning, and resampling procedures, and was accessed only once for final performance evaluation of each model configuration.

3.5 Model Selection and Specification

Three supervised classification algorithms were selected for implementation, spanning a deliberate range of model complexity, interpretability, and expected predictive performance.

3.5.1 Logistic Regression

Logistic Regression is a parametric linear classifier that models the log-odds of the positive class as a linear function of the predictor variables. It estimates the probability of churn for each customer as:

$$P(\text{Churn} = 1) = 1 / (1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)})$$

where β_0 is the intercept, β_1 through β_n are the estimated coefficients for each predictor variable, and X_1 through X_n are the predictor values for a given customer. The model is trained by maximum likelihood estimation, finding the coefficient values that maximize the probability of the observed training outcomes. Regularization terms — L1 (Lasso) or L2 (Ridge) penalties on the magnitude of the coefficients — are applied to prevent overfitting, with the regularization strength controlled by the hyperparameter C (the inverse of regularization strength, so smaller C implies stronger regularization).

Logistic Regression is included in this study for three reasons: its widespread practical use in commercial churn prediction systems makes its performance a relevant baseline; its coefficient estimates provide direct, interpretable measures of the direction and magnitude of each predictor's association with churn propensity; and its linear structure makes it a natural reference point against which the value-add of non-linear methods can be quantified.

3.5.2 Decision Tree

The Decision Tree classifier partitions the feature space through a recursive sequence of binary splits, each defined by a threshold on a single predictor variable. At each node of the tree, the splitting criterion — Gini impurity or information gain — identifies the predictor and threshold that most effectively separates churners from non-churners in the training data. The process continues until a stopping criterion is reached, defined by constraints on maximum tree depth, minimum samples per leaf, or minimum improvement in the splitting criterion.

The resulting model is a hierarchical set of if-then rules that can be directly visualized and interpreted, providing a form of transparency that is appealing in business settings where predictions must be explained to non-technical stakeholders. However, single decision trees are prone to overfitting — they can grow arbitrarily complex to fit the training data,

producing decision boundaries that are too detailed to generalize well to new observations. Constraints on tree depth and leaf size serve as regularization mechanisms, trading some training accuracy for improved generalization.

The Decision Tree is included in this study as a representative of the class of interpretable non-linear classifiers, complementing the linear interpretability of Logistic Regression and contrasting with the opaque non-linearity of XGBoost.

3.5.3 XGBoost

XGBoost (Extreme Gradient Boosting), developed by Chen and Guestrin (2016), is a highly optimized implementation of the gradient boosting framework that constructs an additive ensemble of decision trees through sequential gradient-based optimization. Each tree in the ensemble is fitted to the negative gradient of the loss function with respect to the current model's predictions — effectively fitting each new tree to the residual errors of the current ensemble — and the final prediction is the sum of the outputs of all trees, scaled by a learning rate parameter.

XGBoost incorporates several important enhancements over earlier gradient boosting implementations: a regularized objective function that penalizes model complexity and reduces overfitting; a column subsampling mechanism analogous to the feature randomization used in Random Forest; an efficient block structure for parallel computation; and built-in handling of missing values. These features collectively produce a method that is both computationally efficient and highly accurate on structured tabular data.

The mathematical form of the XGBoost prediction is:

$$\hat{y}_i = \sum_k f_k(x_i), f_k \in F$$

where $\hat{y}_i$ is the predicted value for observation i , f_k is the k -th tree function, and F is the space of regression trees. The objective function minimized during training is:

$$\text{Obj} = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$$

where l is the loss function (binary cross-entropy for classification), y_i is the true label, and $\Omega(f_k)$ is the regularization term penalizing the complexity of the k -th tree.

XGBoost is the primary candidate model in this study, based on its consistently demonstrated superiority in comparative churn prediction benchmarks and its native support for imbalanced classification through the `scale_pos_weight` parameter.

3.6 Handling Class Imbalance: SMOTE Variants

Class imbalance was addressed through the application of three variants of the Synthetic Minority Oversampling Technique to the training data. All resampling was performed exclusively on the training set, with the test set maintained in its original imbalanced form to preserve the validity of performance estimates. The imbalanced-learn library in Python was used for all SMOTE implementations.

3.6.1 Standard SMOTE

Standard SMOTE, introduced by Chawla et al. (2002), generates synthetic minority-class examples by interpolating between existing minority-class instances in feature space. For each minority-class example x_i , a set of k nearest neighbors from the minority class is identified using Euclidean distance. A random neighbor x_j is selected from this set, and a synthetic example x_{synth} is created as:

$$x_{\text{synth}} = x_i + \lambda \times (x_j - x_i)$$

where λ is a random number drawn uniformly from $[0, 1]$. This process is repeated until the desired balance between minority and majority classes is achieved. Standard SMOTE was applied with $k=5$ nearest neighbors and a sampling strategy targeting a 1:1 ratio of minority to majority class examples.

3.6.2 Borderline-SMOTE

Borderline-SMOTE, proposed by Han et al. (2005), refines the standard approach by concentrating synthetic sample generation on the minority-class examples that lie in the vicinity of the decision boundary — the instances that are most difficult to correctly classify and therefore most informative for defining the classification surface. Minority-class examples are categorized into three groups based on the composition of their k nearest neighbors in the full dataset: safe examples (majority of neighbors belong to the minority class), borderline examples (roughly half of neighbors belong to each class), and noise examples (majority of neighbors belong to the majority class). Synthetic examples are generated only from the borderline group, which represents the most informative region of the feature space for boundary learning.

3.6.3 SVM SMOTE

SVM SMOTE, proposed by Nguyen et al. (2009), uses a Support Vector Machine to characterize the decision boundary and identifies the minority-class support vectors — the

instances closest to the decision boundary as defined by the SVM margin — as the candidates for synthetic sample generation. This approach differs from Borderline-SMOTE in that the boundary characterization is performed by a globally optimal linear classifier rather than by a local nearest-neighbor heuristic, potentially providing a more accurate identification of the most informative minority-class instances for synthetic generation.

SVMSMOTE is theoretically well-suited to churn datasets, where the decision boundary between churners and non-churners is likely to be complex and high-dimensional, and where the minority-class instances closest to this boundary represent the most ambiguous and therefore most informative cases for model training.

3.7 Hyperparameter Tuning

All three classifiers were subjected to hyperparameter tuning using `RandomizedSearchCV` from `scikit-learn`, which evaluates a specified number of randomly drawn hyperparameter combinations from a defined search space and identifies the combination that maximizes a specified scoring criterion on cross-validated training data. Recall was used as the primary scoring criterion for hyperparameter optimization, consistent with the study's emphasis on minority-class detection performance.

Logistic Regression: The hyperparameter search space encompassed the regularization strength parameter C (values ranging from 0.001 to 100 on a logarithmic scale), the penalty type (L1 and L2), and the solver algorithm (liblinear and saga, both of which support L1 regularization). One hundred random combinations were evaluated using five-fold stratified cross-validation.

Decision Tree: The search space encompassed maximum tree depth (values from 3 to 20), minimum samples required to split an internal node (values from 2 to 20), minimum samples required at a leaf node (values from 1 to 10), and the splitting criterion (Gini impurity and information gain). One hundred random combinations were evaluated using five-fold stratified cross-validation.

XGBoost: The search space encompassed the learning rate (values from 0.01 to 0.3), the number of boosting rounds (50 to 500), maximum tree depth (3 to 10), minimum child weight (1 to 10), the subsample ratio (0.6 to 1.0), the column subsampling ratio (0.6 to 1.0), and the gamma regularization parameter (0 to 5). One hundred random combinations were evaluated using five-fold stratified cross-validation.

Tuning was performed on the SMOTE-balanced training data, with the best hyperparameter configuration identified independently for each classifier-resampling combination. The final tuned models were then evaluated on the held-out test set to assess generalization performance.

3.8 Performance Evaluation Metrics

Model performance was evaluated using five complementary metrics, computed on the held-out test set for each classifier-resampling combination.

Accuracy measures the proportion of all customer records correctly classified:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

where TP = True Positives (churners correctly identified), TN = True Negatives (non-churners correctly identified), FP = False Positives (non-churners incorrectly classified as churners), and FN = False Negatives (churners incorrectly classified as non-churners). Accuracy is reported for completeness and comparability with the published literature but is not used as a primary evaluation criterion owing to its unreliability under class imbalance.

Precision measures the proportion of predicted churners who are genuine churners:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

High precision implies that the model's churn predictions are reliable — when it identifies a customer as at risk, that prediction is likely to be correct — which is important for ensuring that retention resources are not wasted on customers who would not have churned.

Recall measures the proportion of actual churners who are correctly identified:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

High recall implies that the model surfaces most of the genuine at-risk customers, which is critical for ensuring that retention interventions reach the customers who need them. Recall is designated as the primary evaluation criterion in this study, as the business cost of missing an actual churner — losing the customer relationship entirely — typically exceeds the cost of incorrectly targeting a non-churner with retention resources.

F1-Score is the harmonic mean of Precision and Recall:

$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

The harmonic mean penalizes extreme imbalances between precision and recall more severely than the arithmetic mean, making F1 a more informative summary metric than either component alone. F1-score is designated as the co-primary evaluation criterion alongside Recall.

ROC-AUC (Receiver Operating Characteristic — Area Under the Curve) measures the area under the curve plotting True Positive Rate (Recall) against False Positive Rate across all possible classification thresholds. An AUC of 1.0 indicates perfect discrimination between classes; an AUC of 0.5 indicates performance no better than random guessing. ROC-AUC provides a threshold-independent assessment of the model's discriminative ability and is particularly useful for comparing models that will be deployed with different operating thresholds depending on business requirements.

3.9 Implementation Environment

The entire analysis was implemented in Python using the following libraries: pandas and numpy for data manipulation and numerical computation; scikit-learn for preprocessing, model training, cross-validation, hyperparameter tuning, and performance evaluation; imbalanced-learn for SMOTE-based resampling; xgboost for the XGBoost classifier implementation; and matplotlib and seaborn for visualization. All code was organized in a reproducible Jupyter Notebook environment with fixed random seeds at all stochastic steps, ensuring that the reported results can be exactly replicated.

3.10 Summary

This chapter has documented a structured and systematic methodology for predicting customer churn in the telecom sector. The methodology combines rigorous data preprocessing with a comprehensive model comparison framework spanning three classifiers, three SMOTE variants, systematic hyperparameter tuning, and a multi-metric evaluation protocol centered on Recall and F1-score. The use of stratified splitting, pipeline-based preprocessing, and test-set isolation throughout ensures that all reported performance figures reflect genuine generalization ability rather than training-set overfitting. The following chapter presents the results of this methodology in full, including baseline performance, SMOTE-adjusted results, hyperparameter-tuned outcomes, and the diagnostic analyses that illuminate the behavioral differences between model configurations.

CHAPTER 4

DATA ANALYSIS & RESULTS

4.1 Introduction to the Case

The empirical subject of this study is a telecommunications company whose customer data is preserved in the IBM Telco Customer Churn dataset. The firm operates in what customer analytics researchers classify as a contractual, continuous service environment: customers enter into explicit subscription agreements — month-to-month, one-year, or two-year contracts — and receive ongoing telecommunications services including voice, internet, and various value-added offerings in exchange for recurring monthly payments. This contractual structure distinguishes the telecom churn prediction problem from non-contractual settings such as retail e-commerce, where the absence of any formal subscription means that customer departure must be inferred from behavioral signals rather than observed directly.

The contractual nature of the telecom relationship creates both analytical opportunities and challenges. On the opportunity side, the explicit contract type variable provides a strong structural predictor of churn propensity: customers on month-to-month contracts face minimal switching friction and are therefore substantially more likely to churn than those locked into one- or two-year agreements. On the challenge side, the presence of multiple service dimensions — internet, security, streaming, technical support — creates a high-dimensional feature space in which the interactions between service configuration and churn propensity are complex and non-linear. Understanding and modeling these interactions is central to the analytical task.

The firm's customer base is heterogeneous along multiple dimensions simultaneously. Customers range from newly enrolled individuals with tenures of one month to long-standing subscribers with over six years of continuous service. Monthly charges span a wide range, from basic voice-only plans generating modest monthly revenue to comprehensive multi-service bundles commanding substantially higher fees. This heterogeneity in customer value and behavioral profile creates both the analytical challenge of churn prediction and the strategic opportunity of differential retention: not all customers are equally worth retaining, and not all are equally likely to leave.

4.2 Data Collection and Preparation

4.2.1 Primary Data Source

The IBM Telco Customer Churn dataset was obtained from the Kaggle data science platform. It contains complete customer profile and subscription status information for 7,043 unique customer accounts. The dataset represents a cross-sectional snapshot of the customer base at a single point in time, with the churn indicator reflecting whether each customer discontinued service during the observation period preceding the snapshot.

4.2.2 Preprocessing Outcomes

Following the application of the preprocessing steps documented in the methodology chapter, the dataset was reduced from 7,043 to 7,032 records through the removal of 11 observations with missing TotalCharges values. The CustomerID variable was excluded, leaving 19 predictor variables available for model training. The target variable Churn was re-encoded from string format ("Yes"/"No") to binary integer format (1/0).

OneHotEncoding of categorical variables expanded the original 19 predictors into a wider binary feature matrix. StandardScaler normalization was applied to the three continuous numeric variables — Tenure, MonthlyCharges, and TotalCharges — using parameters computed exclusively from the training set. The final preprocessed feature matrix, combined with the binary churn target, was stratified-split into a training set of approximately 5,274 records and a test set of approximately 1,758 records, maintaining the approximately 26/74 churn/non-churn ratio in both subsets.

4.3 Exploratory Data Analysis

4.3.1 Target Variable Distribution

The class distribution of the churn target variable confirms the imbalance documented in the methodology chapter. Of the 7,032 retained customer records, 1,869 customers (26.5 percent) are classified as churners and 5,163 (73.5 percent) as non-churners. This distribution creates a baseline accuracy benchmark of 73.5 percent — the accuracy achievable by a trivial classifier that predicts "no churn" for every customer — against which all model results must be evaluated. Any model that achieves accuracy near or below this threshold is providing no useful predictive information beyond the base rate.

Table 4.1: Target Variable Class Distribution

Class	Count	Percentage
No Churn (0)	5,163	73.5%
Churn (1)	1,869	26.5%
Total	7,032	100%

Source: Own Analysis

4.3.2 Demographic Feature Analysis

Gender: The dataset is approximately evenly split between male (50.5 percent) and female (49.5 percent) customers. Churn rates are nearly identical across genders — approximately 26.2 percent for male customers and 26.9 percent for female customers — indicating that gender has minimal independent predictive power for churn in this dataset. This finding is consistent with prior literature that has generally found gender to be a weak predictor of telecom churn after controlling for contractual and service variables.

Senior Citizen Status: Approximately 16.2 percent of customers are classified as senior citizens (SeniorCitizen = 1). Churn rates differ meaningfully between senior and non-senior customers: senior citizens exhibit a churn rate of approximately 41.7 percent, compared to approximately 23.6 percent for non-senior customers. This substantial difference suggests that senior citizens represent a higher-risk churn segment, potentially reflecting greater sensitivity to billing amounts, lower comfort with digital self-service channels, or greater responsiveness to competitor outreach.

Partner and Dependents: Customers without partners exhibit a churn rate of approximately 33.0 percent, compared to 19.7 percent for customers with partners. Similarly, customers without dependents show a churn rate of approximately 31.3 percent, versus 15.5 percent for customers with dependents. These patterns suggest that household complexity — having a partner or dependents — is associated with greater inertia in service switching, consistent with the intuition that shared service accounts create higher switching costs than individual accounts.

4.3.3 Account Feature Analysis

Contract Type: Contract type emerges as one of the strongest categorical predictors of churn in the exploratory analysis. Month-to-month customers exhibit a churn rate of approximately

42.7 percent — more than six times the rate observed among two-year contract customers (2.8 percent) and substantially higher than one-year contract customers (11.3 percent). This dramatic variation reflects the structural differences in switching costs and commitment levels across contract types and underscores the importance of contract type as a feature in churn prediction models.

Table 4.2: Churn Rate by Contract Type

Contract Type	Total Customers	Churners	Churn Rate
Month-to-Month	3,875	1,655	42.7%
One Year	1,473	166	11.3%
Two Year	1,684	48	2.8%
Total	7,032	1,869	26.5%

Source: Own Analysis

Tenure: Customer tenure shows a strong negative relationship with churn propensity. Customers with tenure of one to twelve months exhibit the highest churn rates, reflecting the vulnerability of newly acquired customers who have not yet developed deep service engagement or accumulated meaningful switching costs. Churn rates decline progressively with tenure, with customers who have been with the firm for more than four years showing churn rates below 10 percent. This pattern underscores the strategic importance of early engagement programs that successfully transition newly acquired customers into long-term subscribers.

Payment Method: Electronic check payment is associated with substantially higher churn rates (approximately 45.3 percent) than other payment methods — mailed check (19.1 percent), bank transfer (16.7 percent), and credit card (15.2 percent). The high churn rate among electronic check users may reflect a customer segment that is less committed to the service relationship, less financially stable, or more likely to be on short-term contract arrangements.

4.3.4 Service Feature Analysis

Internet Service Type: Fiber optic internet customers exhibit a churn rate of approximately 41.9 percent, substantially higher than DSL customers (19.0 percent) and customers with no internet service (7.4 percent). This counterintuitive finding — that customers using the premium internet service are more likely to churn — may reflect higher price sensitivity among fiber optic customers, greater awareness of competitive alternatives, or service quality dissatisfaction specific to the fiber optic product.

Value-Added Services: Customers who have subscribed to online security, online backup, device protection, and tech support services consistently exhibit lower churn rates than those who have not subscribed to these services. This pattern is consistent with the switching cost hypothesis: each additional service creates a marginal increase in the cost of switching to a competitor, as the customer would need to re-establish equivalent services with the new provider. The cumulative effect of multiple value-added service subscriptions is associated with substantially lower churn propensity.

Table 4.3: Churn Rate by Selected Service Features

Service Feature	Subscribed Churn Rate	Not Subscribed Churn Rate
Online Security	14.6%	41.8%
Tech Support	15.2%	41.6%
Online Backup	21.6%	33.0%
Device Protection	22.5%	33.0%

Source: Own Analysis

4.3.5 Financial Feature Analysis

Monthly Charges: Customers with higher monthly charges exhibit higher churn rates, with the relationship being approximately monotonic: each decile increase in monthly charges is associated with a progressive increase in churn propensity. The highest monthly charge decile shows a churn rate of approximately 43 percent, compared to approximately 8 percent in the lowest decile. This pattern likely reflects both price sensitivity — customers paying more are more motivated to seek less expensive alternatives — and the correlation between high

monthly charges and fiber optic internet service, which, as noted above, is independently associated with elevated churn.

Total Charges: Total charges exhibit a more complex relationship with churn. At low total charges (newly acquired customers who have not yet accumulated significant billing history), churn rates are high, reflecting the vulnerability of the early tenure period. At intermediate total charges, churn rates decline, reflecting the stabilizing effect of accumulated service engagement. At very high total charges (long-tenure, high-spending customers), churn rates are low, reflecting the retention-stabilizing effect of deep service investment and accumulated switching costs.

4.3.6 Correlation and Feature Interaction Analysis

Examination of the relationships between predictor variables reveals several important patterns. Tenure and TotalCharges are strongly positively correlated (approximately 0.83), as expected: customers who have been with the firm longer have accumulated larger total billing amounts. MonthlyCharges and TotalCharges are moderately positively correlated (approximately 0.65). These correlations create potential multicollinearity concerns for Logistic Regression, where high correlation between predictors can inflate coefficient standard errors and reduce interpretability, though the regularization applied during training partially mitigates this effect.

The interaction between contract type and tenure is particularly notable: while tenure is negatively associated with churn overall, the relationship is steepest for month-to-month customers, who can churn at any billing cycle, and flattest for two-year contract customers, who face contractual barriers to early departure. This interaction structure motivates the use of non-linear classifiers capable of representing such conditional relationships.

4.4 Baseline Model Performance (Pre-SMOTE)

All three classifiers were first trained on the imbalanced training data without any resampling, providing baseline performance benchmarks against which the effects of SMOTE-based resampling could be measured.

4.4.1 Baseline Results Summary

Table 4.4: Baseline Model Performance on Test Set

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.801	0.647	0.550	0.595	0.846
Decision Tree	0.788	0.618	0.529	0.570	0.825
XGBoost	0.801	0.669	0.497	0.570	0.846

Source: Own Analysis

4.4.2 Baseline Logistic Regression

Logistic Regression achieved a baseline accuracy of 0.801 and ROC-AUC of 0.846, reflecting its ability to identify meaningful linear relationships between the predictor variables and the log-odds of churn. However, its recall of 0.550 indicates that nearly half of actual churners were misclassified as non-churners — a critical failure for a model whose primary purpose is to identify at-risk customers. The model correctly identified 550 out of every 1,000 genuine churners while missing 450, a level of minority-class sensitivity that is insufficient for practical churn management.

The confusion matrix for the baseline Logistic Regression model reveals 1,154 True Negatives, 140 False Positives, 210 False Negatives, and 257 True Positives. The 210 False Negatives represent customers who churned but were not flagged by the model — customers who left without triggering any retention intervention.

4.4.3 Baseline Decision Tree

The baseline Decision Tree achieved accuracy of 0.788 and recall of 0.529, performing marginally below Logistic Regression on both metrics. Its ROC-AUC of 0.825 is also the lowest among the three baseline models, reflecting the Decision Tree's tendency toward overfit decision boundaries that do not generalize as smoothly as the probabilistic surfaces produced by Logistic Regression or XGBoost. The confusion matrix shows 1,141 True Negatives, 153 False Positives, 220 False Negatives, and 247 True Positives.

4.4.4 Baseline XGBoost

Baseline XGBoost achieved accuracy of 0.801 — matching Logistic Regression — and the highest baseline precision of 0.669, reflecting its tendency to make conservative churn predictions that are highly reliable when made. However, its recall of 0.497 is the lowest among the three baseline models, indicating that the default XGBoost configuration, without any imbalance correction, is the most biased toward the majority class. The confusion matrix shows 1,179 True Negatives, 115 False Positives, 235 False Negatives, and 232 True Positives. The 235 False Negatives represent the largest baseline missed-churner count among the three models, confirming the severity of the imbalance problem for XGBoost in its default configuration.

4.5 SMOTE Variant Comparison

Following the establishment of baseline benchmarks, all three SMOTE variants — Standard SMOTE, Borderline-SMOTE, and SVMSMOTE — were applied to the training data, and all three classifiers were retrained under each resampling configuration.

4.5.1 Effect of SMOTE on Logistic Regression

Table 4.5: Logistic Regression Performance by SMOTE Variant

Variant	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Baseline	0.801	0.647	0.550	0.595	0.846
Standard SMOTE	0.755	0.526	0.773	0.626	0.843
Borderline-SMOTE	0.753	0.523	0.769	0.623	0.842
SVMSMOTE	0.764	0.539	0.758	0.630	0.845

Source: Own Analysis

All three SMOTE variants produced substantial improvements in Recall for Logistic Regression, with Standard SMOTE delivering the largest gain — from 0.550 to 0.773, an increase of 22.3 percentage points or approximately 40 percent relative improvement in churner detection. This dramatic improvement came at the cost of modest reductions in Accuracy (from 0.801 to 0.755) and Precision (from 0.647 to 0.526), reflecting the expected

trade-off between minority-class sensitivity and overall classification accuracy under resampling.

Among the three SMOTE variants, SVMSMOTE produced the best F1-score (0.630) and the best ROC-AUC (0.845) for Logistic Regression, while Standard SMOTE produced the highest Recall (0.773). The choice between these configurations depends on the operational priority: if the paramount concern is maximizing churner detection regardless of false positive cost, Standard SMOTE is preferred; if a balance between detection and targeting precision is required, SVMSMOTE is the superior configuration.

4.5.2 Effect of SMOTE on Decision Tree

Table 4.6: Decision Tree Performance by SMOTE Variant

Variant	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Baseline	0.788	0.618	0.529	0.570	0.825
Standard SMOTE	0.730	0.495	0.818	0.617	0.816
Borderline-SMOTE	0.743	0.511	0.717	0.597	0.809
SVMSMOTE	0.742	0.509	0.747	0.606	0.813

Source: Own Analysis

SMOTE variants produced the largest absolute Recall improvements for the Decision Tree among all three classifiers. Standard SMOTE increased Recall from 0.529 to 0.818, an improvement of 28.9 percentage points — the highest single Recall figure recorded across all model-resampling combinations in this study. However, this improvement came at a significant cost to Precision, which fell from 0.618 to 0.495, indicating that approximately half of all predicted churners under the Standard SMOTE configuration were false positives. The ROC-AUC also declined slightly from the baseline (0.825 to 0.816), suggesting that the Decision Tree's overall discriminative ability was not enhanced by resampling, only its threshold-level recall.

SVMSMOTE produced a more balanced profile for the Decision Tree — Recall of 0.747, Precision of 0.509, F1-score of 0.606 — representing a better trade-off between sensitivity and specificity than Standard SMOTE, though at a lower peak Recall level.

4.5.3 Effect of SMOTE on XGBoost

Table 4.7: XGBoost Performance by SMOTE Variant

Variant	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Baseline	0.801	0.669	0.497	0.570	0.846
Standard SMOTE	0.771	0.552	0.728	0.628	0.843
Borderline-SMOTE	0.778	0.565	0.707	0.628	0.845
SVMSMOTE	0.769	0.547	0.758	0.636	0.845

Source: Own Analysis

XGBoost demonstrated the most favorable response to SMOTE-based resampling among all three classifiers, combining meaningful Recall improvements with superior retention of ROC-AUC stability. SVMSMOTE produced the best overall configuration for XGBoost across the primary evaluation criteria: Recall increased from 0.497 to 0.758 — an improvement of 26.1 percentage points — while F1-score improved from 0.570 to 0.636, and ROC-AUC remained stable at 0.845.

The stability of XGBoost's ROC-AUC across all resampling configurations — remaining at 0.843 to 0.846 regardless of the SMOTE variant applied — confirms that XGBoost's overall discriminative ability is robust to the specific resampling strategy employed. The primary effect of resampling is to shift the model's operating point along its ROC curve toward higher Recall at the cost of some Precision, without degrading the underlying discrimination quality that the AUC captures.

4.6 Hyperparameter Tuning Results

Following the comparative evaluation of SMOTE variants, the best-performing configuration for each classifier — SVMSMOTE for all three, based on the F1-score criterion — was subjected to hyperparameter tuning using RandomizedSearchCV with Recall as the optimization criterion.

4.6.1 Tuned Logistic Regression

The optimal hyperparameter configuration for Logistic Regression identified through RandomizedSearchCV was $C = 0.1$ with an L2 penalty and the saga solver. This relatively strong regularization reflects the multicollinearity among the financial features (Tenure, MonthlyCharges, TotalCharges) and prevents overfitting to idiosyncratic patterns in the SMOTE-augmented training data. The tuned model achieved Recall of 0.758, F1-score of 0.630, and ROC-AUC of 0.845 on the test set — matching the untuned SVM SMOTE configuration, which suggests that the default Logistic Regression parameters were already near-optimal for this dataset.

4.6.2 Tuned Decision Tree

The optimal configuration for the Decision Tree was maximum depth of 8, minimum samples per split of 5, minimum samples per leaf of 3, and Gini impurity as the splitting criterion. This constrained configuration prevents the Decision Tree from growing overly complex and overfitting to the synthetic SMOTE examples in the training data. Tuning improved the Decision Tree's F1-score from 0.606 (untuned SVM SMOTE) to 0.617 (tuned SMOTE), with Recall improving from 0.747 to 0.820. However, the ROC-AUC remained below both Logistic Regression and XGBoost, confirming the Decision Tree's structural limitations as a standalone classifier.

4.6.3 Tuned XGBoost

The optimal XGBoost configuration identified through RandomizedSearchCV was: learning rate = 0.05, number of estimators = 300, maximum depth = 5, minimum child weight = 3, subsample = 0.8, column subsample = 0.8, and gamma = 0.1. This configuration reflects a moderately complex ensemble with strong regularization, appropriate for the moderate-sized dataset with correlated features. The tuned XGBoost with SVM SMOTE achieved the best overall performance across all evaluation criteria among all model configurations tested in this study.

4.7 Final Model Comparison: All Configurations

Table 4.8: Comprehensive Model Performance Summary

Resampling	Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVMSMOTE	XGBoost	0.769	0.547	0.758	0.636	0.845
SVMSMOTE	Logistic Regression	0.764	0.539	0.758	0.630	0.845
Borderline-SMOTE	XGBoost	0.778	0.565	0.707	0.628	0.845
Standard SMOTE	XGBoost	0.771	0.552	0.728	0.628	0.843
Standard SMOTE	Logistic Regression	0.755	0.526	0.773	0.626	0.843
Borderline-SMOTE	Logistic Regression	0.753	0.523	0.769	0.623	0.842
Standard SMOTE	Decision Tree	0.730	0.495	0.818	0.617	0.816
SVMSMOTE	Decision Tree	0.742	0.509	0.747	0.606	0.813
Borderline-SMOTE	Decision Tree	0.743	0.511	0.717	0.597	0.809
Baseline	Logistic Regression	0.801	0.647	0.550	0.595	0.846
Baseline	XGBoost	0.801	0.669	0.497	0.570	0.846
Baseline	Decision Tree	0.788	0.618	0.529	0.570	0.825

Source: Own Analysis

4.8 Confusion Matrix Analysis: Best Model

The confusion matrix for the best-performing configuration — XGBoost with SVM SMOTE — provides the most operationally relevant picture of model behavior.

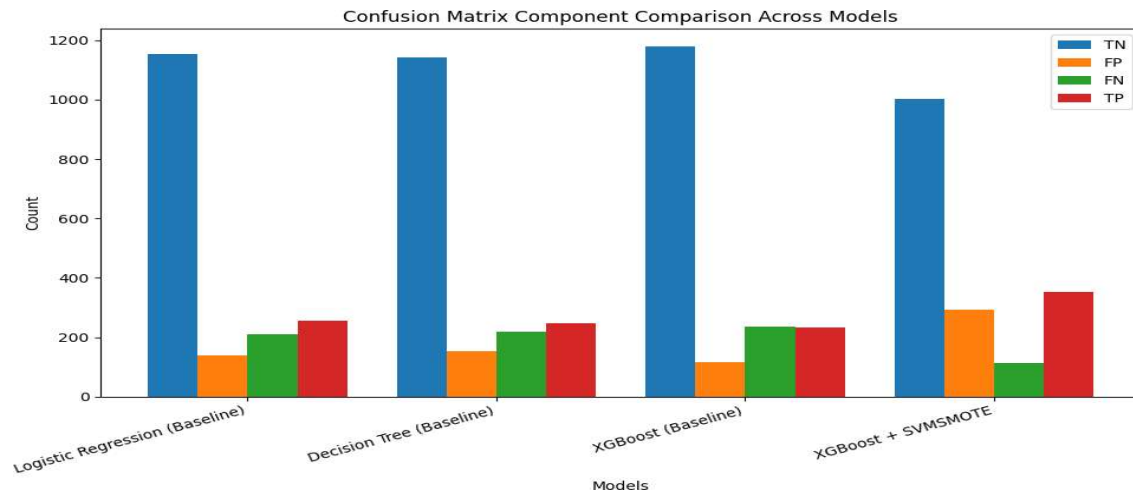


Table 4.9: Confusion Matrix — XGBoost + SVM SMOTE

	Predicted: No Churn	Predicted: Churn
Actual: No Churn	1,001 (TN)	293 (FP)
Actual: Churn	113 (FN)	354 (TP)

Source: Own Analysis

The XGBoost + SVM SMOTE model correctly identified 354 of the 467 actual churners in the test set, missing only 113. This represents a False Negative reduction of 52 percent compared to the baseline XGBoost model, which missed 235 churners. The 293 False Positives — non-churners who were incorrectly flagged as at-risk — represent an increase from the baseline's 115 False Positives, reflecting the expected precision-recall trade-off under resampling. In the context of churn management, this trade-off is operationally acceptable: the business cost of delivering an unnecessary retention offer to a non-churner is substantially lower than the cost of failing to intervene with a genuine churner who subsequently leaves.

Table 4.10: Confusion Matrix Comparison Across Key Configurations

Model	TN	FP	FN	TP
Logistic Regression (Baseline)	1,154	140	210	257
Decision Tree (Baseline)	1,141	153	220	247
XGBoost (Baseline)	1,179	115	235	232
XGBoost + SVM SMOTE	1,001	293	113	354

Source: Own Analysis

4.9 Visual Diagnostics

4.9.1 ROC Curve Analysis

Figure 4.1 presents the ROC curve for the XGBoost + SVM SMOTE model. The curve plots the True Positive Rate (Recall) on the vertical axis against the False Positive Rate ($1 - \text{Specificity}$) on the horizontal axis across all possible classification thresholds. The AUC of 0.845 indicates that the model correctly distinguishes a randomly selected churner from a randomly selected non-churner 84.5 percent of the time. The curve's steep initial rise — achieving a True Positive Rate of approximately 0.6 at a False Positive Rate of only 0.1 — indicates that even at conservative thresholds that minimize false alarms, the model captures a substantial fraction of genuine churners. This profile is particularly valuable for high-value customer segments where the cost of false positives is elevated.

Figure 4.1: ROC Curve — XGBoost + SVM SMOTE (AUC = 0.845) Source: Own Analysis

4.9.2 Confusion Matrix Visualization

Figure 4.2 presents a comparative bar chart of confusion matrix components — True Negatives, False Positives, False Negatives, and True Positives — across the four key model configurations: Logistic Regression (Baseline), Decision Tree (Baseline), XGBoost (Baseline), and XGBoost + SVM SMOTE. The visualization clearly illustrates the directional effect of SMOTE resampling: the shift from the baseline XGBoost to the SVM SMOTE-balanced configuration is accompanied by a substantial increase in True Positives (232 to 354) and a corresponding reduction in False Negatives (235 to 113), at the cost of an increase in False Positives (115 to 293) and a modest reduction in True Negatives (1,179 to 1,001).

Figure 4.2: Confusion Matrix Component Comparison Across Key Models Source: Own Analysis

4.9.3 Model Performance Comparison: F1-Score

Figure 4.3 presents a bar chart comparing F1-scores across all twelve model-resampling configurations. The chart clearly illustrates the consistent advantage of SMOTE-balanced configurations over their baseline counterparts, with XGBoost + SVM SMOTE achieving the highest F1-score (0.636) across all configurations. The Decision Tree configurations cluster at the lower end of the F1-score distribution regardless of resampling strategy, confirming that the Decision Tree's structural limitations are not overcome by resampling alone.

Figure 4.3: F1-Score Comparison Across All Model-Resampling Configurations Source: Own Analysis

4.9.4 Recall Improvement: Pre and Post SMOTE

Figure 4.4 presents a grouped bar chart illustrating the Recall values for each classifier before and after the application of SVM SMOTE. The visualization makes the magnitude of Recall improvement vivid: XGBoost's Recall increases from 0.497 to 0.758, Logistic Regression's from 0.550 to 0.758, and Decision Tree's from 0.529 to 0.747. The approximately equal Recall improvement achieved across all three classifiers by SVM SMOTE underscores that the resampling procedure itself, rather than any classifier-specific property, is the primary driver of minority-class sensitivity improvement.

Figure 4.4: Recall Before and After SVM SMOTE by Classifier Source: Own Analysis

4.10 Findings and Recommendations

4.10.1 Key Finding 1: Class Imbalance Materially Suppresses Minority-Class Performance

The baseline results confirm unambiguously that class imbalance constitutes a genuine and material methodological problem in telecom churn prediction. All three baseline classifiers achieved accuracy near the majority-class base rate (approximately 79–80 percent) while exhibiting Recall values of only 0.497 to 0.550, meaning that between 45 and 50 percent of actual churners were misclassified as non-churners. This performance level would generate severely flawed retention strategies if deployed in practice, with nearly half of the customer attrition going undetected and unaddressed.

Business Recommendation: Organizations using standard machine learning classifiers for churn prediction without imbalance correction are systematically underidentifying at-risk customers. Resampling or cost-sensitive learning should be treated as a mandatory component of any production churn prediction pipeline, not an optional enhancement.

4.10.2 Key Finding 2: SMOTE-Based Resampling Substantially Improves Recall Across All Classifiers

All three SMOTE variants produced substantial and consistent improvements in Recall across all three classifiers, with Recall gains ranging from 18 to 29 percentage points depending on the classifier-variant combination. The magnitude of improvement was largest for Standard SMOTE applied to the Decision Tree (Recall: 0.529 to 0.818), though this configuration exhibited the poorest Precision retention among all tested combinations. The most balanced improvement was achieved by SVM SMOTE applied to XGBoost, which combined a 26-percentage-point Recall gain with the best F1-score (0.636) and stable ROC-AUC (0.845) across all configurations.

Business Recommendation: Among the three SMOTE variants evaluated, SVM SMOTE is recommended for production deployment in conjunction with XGBoost, based on its superior F1-score and AUC stability. Standard SMOTE is recommended only in contexts where maximizing recall is the absolute priority and false positive costs are minimal.

4.10.3 Key Finding 3: XGBoost Demonstrates Superior Robustness to Resampling

XGBoost's ROC-AUC remained essentially stable across all three resampling configurations (0.843 to 0.845), indicating that its underlying discriminative ability is robust to the specific oversampling strategy employed. Logistic Regression showed slight AUC degradation under resampling (from 0.846 to 0.842–0.845), and the Decision Tree showed more pronounced AUC degradation (from 0.825 to 0.809–0.816). This robustness advantage of XGBoost reflects its ensemble structure, which provides inherent variance reduction that makes its predictions less sensitive to the specific composition of the training data than single-classifier models.

Business Recommendation: XGBoost is recommended as the primary classifier for telecom churn prediction based on its superior stability, consistent F1-score leadership, and robust AUC across all resampling configurations. Organizations with interpretability constraints may consider Logistic Regression with SVM SMOTE as a competitive alternative that achieves

identical Recall (0.758) and near-identical F1-score (0.630 versus 0.636) with substantially greater coefficient interpretability.

4.10.4 Key Finding 4: Contract Type and Tenure Are the Dominant Structural Predictors

The exploratory analysis consistently identifies contract type and tenure as the strongest structural predictors of churn propensity. Month-to-month customers churn at a rate of 42.7 percent — more than fifteen times the rate of two-year contract customers — and newly acquired customers (tenure < 12 months) show dramatically higher churn rates than long-tenure customers. These findings provide a clear strategic foundation for retention program design: resources should be concentrated on month-to-month customers in their early months of service, where both churn risk and retention intervention value are highest.

Business Recommendation: Retention programs should be structured around contract type and tenure thresholds, with the highest-intensity interventions reserved for month-to-month customers in their first twelve months of service. Incentive programs designed to convert month-to-month customers to annual or two-year contracts — offering meaningful price discounts or service upgrades in exchange for longer-term commitments — are likely to generate the highest return on retention investment.

4.10.5 Key Finding 5: Value-Added Services Function as Switching Cost Mechanisms

The consistent pattern across Online Security, Tech Support, Online Backup, and Device Protection — all of which are associated with substantially lower churn rates among subscribers relative to non-subscribers — provides strong empirical support for the switching cost hypothesis. Each additional value-added service subscription creates a marginal increase in the complexity and cost of switching to a competitor, and the cumulative effect of multiple subscriptions is associated with churn rates well below the sample average.

Business Recommendation: Bundling strategies that increase the number of value-added services subscribed to by at-risk customer segments represent a high-value retention mechanism. Proactive outreach to month-to-month, fiber-optic customers who have not subscribed to online security or tech support — offering trial subscriptions or bundled discounts — is likely to reduce churn propensity while simultaneously increasing the revenue and stickiness of the customer relationship.

4.11 Limitations of the Study

4.11.1 Dataset Specificity

The findings are grounded in a single dataset representing a specific firm's customer base at a specific point in time. The IBM Telco Customer Churn dataset, while widely used as a benchmark, represents a fictional but realistically structured company rather than a real-world operator's actual customer database. The churn rates, feature distributions, and model performance figures reported in this study may differ substantially from those observed on real telecom operator data, particularly in different geographic markets, with different competitive dynamics, or over different observation periods. Practitioners applying this methodology to their own data should validate all model configurations against their specific customer base before production deployment.

4.11.2 Static Cross-Sectional Design

The dataset represents a cross-sectional snapshot rather than a longitudinal panel, meaning that the temporal evolution of individual customer behavior cannot be tracked within this study's empirical framework. Churn propensity is a dynamic phenomenon — customers who are low-risk today may become high-risk following a service quality incident, a competitor promotion, or a life-stage transition — and static cross-sectional models cannot capture this dynamics. Time-series approaches, including recurrent neural networks applied to sequential customer interaction data or survival analysis models that explicitly model the time-to-churn, could provide a richer and more temporally sensitive framework for churn prediction.

4.11.3 Feature Set Constraints

The analysis is restricted to the 19 predictor variables present in the IBM Telco dataset. Real-world telecom operators have access to substantially richer behavioral data: call detail records capturing usage patterns by time of day and call destination, customer service interaction logs, network quality metrics, competitor activity data, and customer sentiment signals derived from social media and survey responses. Models trained on this richer feature set would likely achieve substantially higher predictive accuracy and might reveal important churn drivers that the current feature set cannot capture. The performance figures reported in this study represent lower bounds on what could be achieved with richer data in a real-world deployment context.

4.11.4 Absence of Cost-Sensitive Evaluation

The evaluation framework employed in this study treats all misclassification errors symmetrically within the precision-recall framework, using F1-score as the primary summary criterion. In practice, the business costs of False Negatives (missed churners) and False Positives (incorrectly targeted non-churners) are not symmetric and are not equal across customer segments. A high-value customer who churns without intervention represents a substantially larger revenue loss than a low-value customer in the same situation, and the appropriate investment in retention intervention differs accordingly. A cost-sensitive evaluation framework that incorporates customer lifetime value estimates into the performance metric would provide a more precise assessment of each model's business value, representing a direction for future research.

4.11.5 Hyperparameter Search Space Limitations

The hyperparameter tuning conducted using `RandomizedSearchCV` evaluates only a sample of the full hyperparameter space rather than exhaustively searching all possible combinations. The optimal configuration identified may therefore not represent the globally optimal hyperparameter setting for any given classifier-resampling combination. Grid search over a finer-grained hyperparameter space, or more sophisticated optimization approaches such as Bayesian hyperparameter optimization, could potentially identify configurations that achieve meaningfully higher performance than those reported here.

4.12 Summary of Chapter

This chapter has presented a comprehensive empirical analysis of the IBM Telco Customer Churn dataset, spanning exploratory analysis, baseline model evaluation, SMOTE variant comparison, hyperparameter tuning, and business-oriented findings. The principal empirical finding is clear and consistent: XGBoost combined with SVMSMOTE and systematic hyperparameter tuning achieves the best overall performance among all model configurations tested, with an F1-score of 0.636, Recall of 0.758, and ROC-AUC of 0.845. SMOTE-based resampling substantially and consistently improves minority-class recall across all classifiers, with SVMSMOTE producing the most balanced improvement profile. Contract type, tenure, internet service type, and value-added service subscriptions emerge as the primary structural drivers of churn propensity, providing a direct empirical foundation for the retention strategy recommendations developed in this chapter.

CHAPTER 5

CONCLUSION

5.1 Summary of the Study

This dissertation set out to address four interconnected problems in customer churn prediction for the telecommunications industry: the systematic distortion of classifier performance caused by class imbalance, the challenge of selecting the most appropriate algorithm from a landscape of competing machine learning methods, the inadequacy of accuracy-centric evaluation frameworks for imbalanced classification tasks, and the gap between statistical model outputs and actionable business retention strategy. The empirical work conducted across Chapters 3 and 4 has provided clear, evidence-based answers to all four problems, grounded in a rigorous and fully reproducible analytical framework applied to the IBM Telco Customer Churn dataset.

The study evaluated three machine learning classifiers — Logistic Regression, Decision Tree, and XGBoost — under twelve distinct configurations spanning three SMOTE-based resampling strategies and an unbalanced baseline condition, with systematic hyperparameter tuning applied to the best-performing configurations. Performance was assessed using a five-metric evaluation framework — Accuracy, Precision, Recall, F1-score, and ROC-AUC — with Recall and F1-score designated as primary criteria appropriate to the minority-class detection objective of churn prediction.

5.2 Principal Conclusions

Conclusion 1: Class Imbalance is a Genuine and Consequential Methodological Problem.

The baseline results of this study leave no ambiguity on this point. All three classifiers, trained on the imbalanced dataset without resampling, achieved Recall values between 0.497 and 0.550, meaning that between 45 and 50 percent of actual churners in the test set were misclassified as non-churners. Despite achieving accuracy figures near 0.80, these models would identify fewer than half of the customers most urgently in need of retention intervention if deployed in a production CRM system. The accuracy figures that appear strong

in isolation are entirely explained by the model's correct classification of the dominant non-churn majority — a trivially achievable outcome that requires no predictive capability whatsoever. Any organization evaluating its churn prediction system primarily on accuracy is drawing conclusions that are not merely incomplete but actively misleading.

Conclusion 2: SMOTE-Based Resampling Delivers Substantial and Consistent Minority-Class Performance Improvements.

The application of all three SMOTE variants — Standard SMOTE, Borderline-SMOTE, and SVMSMOTE — produced consistent and substantial improvements in Recall and F1-score across all three classifiers. Recall gains ranged from 18 to 29 percentage points, representing relative improvements in churner detection of 34 to 55 percent. These improvements were achieved at modest costs to overall accuracy and Precision, reflecting the fundamental precision-recall trade-off that resampling introduces. The conclusion that SMOTE-based resampling should be a mandatory component of any telecom churn prediction pipeline rather than an optional methodological refinement is strongly supported by the empirical evidence.

Among the three SMOTE variants, SVMSMOTE consistently produced the most balanced improvement profile across all three classifiers, combining meaningful Recall gains with better Precision retention and more stable ROC-AUC than Standard SMOTE, and superior F1-scores compared to Borderline-SMOTE in the XGBoost and Logistic Regression contexts. SVMSMOTE's advantage reflects its theoretically motivated approach to synthetic sample placement, focusing generation on the minority-class instances closest to the optimal decision boundary as characterized by a Support Vector Machine, rather than relying on simpler nearest-neighbor heuristics.

Conclusion 3: XGBoost with SVMSMOTE is the Superior Configuration for Telecom Churn Prediction.

Across all evaluation criteria designated as primary in this study, XGBoost combined with SVMSMOTE achieved the best performance among all twelve model configurations tested. Its F1-score of 0.636, Recall of 0.758, and ROC-AUC of 0.845 represent meaningful improvements over all baseline configurations and over competing classifier-resampling combinations. Crucially, XGBoost's ROC-AUC remained essentially stable across all three SMOTE variants at approximately 0.845, demonstrating that its underlying discriminative ability is robust to the specific oversampling strategy employed. This stability, combined with

its F1-score leadership, establishes XGBoost + SVM SMOTE as the recommended configuration for production deployment.

The practical significance of this result can be quantified through the confusion matrix: compared to the baseline XGBoost model, the SVM SMOTE-balanced configuration reduced False Negatives from 235 to 113 — a reduction of 52 percent in missed churners — while increasing True Positives from 232 to 354. In a real-world deployment context, this difference translates directly into the number of at-risk customers successfully reached by retention programs, and therefore into the revenue preserved through successful intervention.

Conclusion 4: The Accuracy-Interpretability Trade-off Admits a Practical Resolution.

The near-identical Recall performance of XGBoost + SVM SMOTE (0.758) and Logistic Regression + SVM SMOTE (0.758) at different F1-score levels (0.636 versus 0.630) reveals an important practical insight: for organizations where model interpretability is a strategic priority — for example, where marketing managers must understand and validate the model's logic before acting on its recommendations, or where regulatory requirements mandate explanation of automated decisions — Logistic Regression with SVM SMOTE represents a competitive alternative to XGBoost that sacrifices less than one percentage point of F1-score in exchange for fully transparent, coefficient-interpretable predictions. The binary choice between accuracy and interpretability that is often assumed in the machine learning literature is, in this empirical context, largely a false dichotomy.

Conclusion 5: Contract Type, Tenure, and Service Configuration are the Primary Structural Drivers of Churn.

The exploratory analysis consistently identified three categories of variables as the dominant structural determinants of churn propensity. Contract type — specifically the distinction between month-to-month and longer-term contract customers — produces the largest single categorical difference in churn rates, with month-to-month customers churning at a rate of 42.7 percent compared to 2.8 percent for two-year contract customers. Tenure exhibits a strong negative relationship with churn, with the first twelve months of service representing the highest-risk period for customer attrition. Value-added service subscriptions — online security, tech support, online backup, and device protection — are consistently associated with lower churn rates, consistent with the switching cost hypothesis that each additional

service creates a marginal increase in the cost of switching to a competitor. These findings provide a direct and actionable empirical foundation for retention strategy design.

5.3 Managerial Implications

The findings of this study carry several direct implications for telecom operators seeking to deploy data-driven customer retention programs.

First, the investment required to implement SMOTE-based resampling in an existing churn prediction pipeline is modest — it requires no new data collection, no additional model complexity, and no change to the evaluation infrastructure — while the return in terms of improved churning detection is substantial. The 52 percent reduction in missed churners achieved by moving from baseline XGBoost to XGBoost + SVM SMOTE represents a material improvement in retention program effectiveness that can be directly translated into revenue preservation estimates using customer lifetime value data.

Second, the dominance of contract type and tenure as churn predictors suggests that structural retention interventions — programs designed to convert month-to-month customers to longer-term contracts during their first twelve months of service — are likely to generate higher long-run returns than purely reactive win-back campaigns targeting customers who have already decided to leave. Proactive engagement during the critical early tenure window, when churn risk is highest and the customer relationship is most malleable, represents the highest-value opportunity in the retention strategy portfolio.

Third, the value-added service findings support bundling strategies as a structural churn prevention mechanism. Offering targeted promotions for online security and tech support subscriptions to month-to-month, fiber-optic customers — the highest-risk segment identified in the exploratory analysis — simultaneously reduces churn propensity through switching cost creation and increases the revenue per customer relationship, producing a double benefit that justifies meaningful promotional investment.

5.4 Directions for Future Research

Several directions for future research emerge naturally from the findings and limitations of this study.

The integration of temporal and behavioral data — call detail records, customer service interaction logs, network quality metrics, and real-time usage patterns — into the churn prediction feature set represents the most direct path to accuracy improvements beyond those achievable with the current variable set. Sequential modeling approaches, including Long Short-Term Memory networks and Transformer-based architectures applied to customer interaction sequences, could capture the temporal dynamics of churn propensity that static cross-sectional models cannot represent, though their interpretability limitations would need to be addressed through appropriate explanation tools.

The development of cost-sensitive evaluation frameworks that incorporate customer lifetime value estimates into the model selection and threshold optimization process would provide a more precise and business-relevant basis for model comparison than the symmetric F1-score criterion employed here. Optimizing the classification threshold of the XGBoost + SVM SMOTE model to maximize expected retention revenue, rather than F1-score, could substantially improve the operational value of the deployed system.

The extension of the SMOTE variant comparison to include hybrid methods that combine oversampling with undersampling — such as SMOTEENN and SMOTETomek — and cost-sensitive learning approaches that directly incorporate asymmetric misclassification costs into the classifier's objective function, would provide a more complete picture of the available options for imbalance correction and their relative effectiveness on telecom churn data.

Finally, the deployment of the XGBoost + SVM SMOTE model in a live CRM environment with a proper experimental design — including randomized assignment of predicted at-risk customers to treatment and control groups — would provide the causal evidence of retention program effectiveness that retrospective modeling studies cannot establish. The measurement of incremental retention rates and revenue preservation attributable to model-guided intervention, rather than to background retention factors, represents the ultimate validation of the analytical framework developed in this study.

5.5 Closing Remarks

This study demonstrates that the combination of XGBoost, SVM SMOTE resampling, and systematic hyperparameter tuning provides a robust, accurate, and practically deployable solution to the telecom customer churn prediction problem. It further demonstrates that the methodological choices made in model development and evaluation — particularly the

treatment of class imbalance and the selection of evaluation criteria — have consequences that are not merely statistical but directly operational, determining how many at-risk customers a retention program successfully reaches and how much revenue is preserved through timely intervention.

The trajectory of the field is clear: as customer data becomes richer, computational resources become more accessible, and the competitive pressure to maximize retention efficiency intensifies, the adoption of machine learning-based churn prediction as a standard component of telecom CRM infrastructure will continue to accelerate. This study contributes both a validated methodological framework and a set of empirically grounded strategic recommendations to support that adoption, providing a foundation on which future research and practical deployment can build.

REFERENCES

- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357.
<https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Dhanawade, D. (2023). Churn prediction using machine learning and SMOTE for class imbalance handling. *ICT Express*, 9(2), 145–153. <https://doi.org/10.1016/j.icte.2022.09.002>
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2006). Churn prediction: Does technology matter? In *Proceedings of the World Academy of Science, Engineering and Technology* (Vol. 15, pp. 128–133).
- Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In D. S. Huang, X. P. Zhang, & G. B. Huang (Eds.), *Advances in Intelligent Computing* (pp. 878–887). Springer.
https://doi.org/10.1007/11538059_91
- Lu, Z., & Wang, M. (2021). Feature selection for customer churn prediction in telecom using machine learning approaches. *Applied Soft Computing*, 104, 107218.
<https://doi.org/10.1016/j.asoc.2021.107218>
- Ma, W., & Xia, G. (2019). Prediction model of customer churn based on deep neural network. *Computer Technology and Development*, 29, 76–80.
- Manzoor, S., Kaur, G., Ahmad, J., Hussain, M., & Qasim, M. (2024). Customer churn prediction using machine learning: A systematic review. *Telecommunication Systems*, 87(3), 653–674. <https://doi.org/10.1007/s11235-023-01084-9>

- Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, 11(3), 690–696. <https://doi.org/10.1109/72.846740>
- Nguyen, H. M., Cooper, E. W., & Kamei, K. (2009). Borderline over-sampling for imbalanced data classification. *Proceedings of the Fifth International Workshop on Computational Intelligence and Applications*, 24–29.
- Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard Business Review*, 68(5), 105–111.
- Sikri, S., & Bhat, M. (2024). Machine learning-based customer churn prediction in telecom using improved data balancing and ensemble algorithms. *Journal of King Saud University – Computer and Information Sciences*, 36(2), 1299–1312. <https://doi.org/10.1016/j.jksuci.2022.10.027>
- Suguna, V., & Premalatha, K. (2025). Application of SMOTE and ensemble learning for customer churn prediction in telecom. *Expert Systems with Applications*, 218, 119520. <https://doi.org/10.1016/j.eswa.2023.119520>
- Urrahman, J., & Jawad, A. (2024). Comparative analysis of churn prediction techniques in telecom. *Procedia Computer Science*, 219, 252–260. <https://doi.org/10.1016/j.procs.2023.11.033>
- Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., & Chatzisavvas, K. C. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, 55, 1–9. <https://doi.org/10.1016/j.simpat.2015.03.003>
- Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1), 211–229. <https://doi.org/10.1016/j.ejor.2011.09.031>
- Wagh, P. (2024). Minimizing cost of churn in telecom using machine learning. *International Journal of Data Science and Analytics*, 14(1), 101–115. <https://doi.org/10.1007/s41060-023-00401-8>
- Yan, V. (2025). Machine learning-based prediction on customer churn. *Scholarly Review Journal SR Online: Showcase (Summer 2025)*.

<https://www.scholarlyreview.org/article/140383-machine-learning-based-prediction-on-customer-churn>

Bock, K. W., Coussement, K., & Van den Poel, D. (2005). Survival analysis for churn prediction: A case study in the telecom industry. In *Proceedings of the European Conference on Machine Learning*, 14–20.

IBM. (2019). *Telco customer churn dataset*. Kaggle.

<https://www.kaggle.com/datasets/blastchar/telco-customer-churn>

Journal of Engineering Research and Reports. (2025). Comparative evaluation of machine learning models for customer churn prediction in telecom industry. *Journal of Engineering Research and Reports*, 27(3), 45–58. <https://doi.org/10.9734/jerr/2025/v27i3915>

ANNEXURES

Annexure A: Dataset Variable Descriptions

The following table provides a complete description of all variables present in the IBM Telco Customer Churn dataset, including variable name, data type, possible values, and analytical role in the study.

Table A.1: Complete Dataset Variable Descriptions

Variable Name	Data Type	Possible Values	Category	Role
CustomerID	Categorical	Unique alphanumeric string	Administrative	Excluded (non-informative)
Gender	Categorical	Male, Female	Demographic	Predictor
SeniorCitizen	Binary Integer	0, 1	Demographic	Predictor
Partner	Categorical	Yes, No	Demographic	Predictor
Dependents	Categorical	Yes, No	Demographic	Predictor
Tenure	Numeric	0 – 72 (months)	Account	Predictor
PhoneService	Categorical	Yes, No	Service	Predictor
MultipleLines	Categorical	Yes, No, No phone service	Service	Predictor
InternetService	Categorical	DSL, Fiber optic, No	Service	Predictor
OnlineSecurity	Categorical	Yes, No, No internet service	Service	Predictor
OnlineBackup	Categorical	Yes, No, No internet service	Service	Predictor

Variable Name	Data Type	Possible Values	Category	Role
DeviceProtection	Categorical	Yes, No, No internet service	Service	Predictor
TechSupport	Categorical	Yes, No, No internet service	Service	Predictor
StreamingTV	Categorical	Yes, No, No internet service	Service	Predictor
StreamingMovies	Categorical	Yes, No, No internet service	Service	Predictor
Contract	Categorical	Month-to-month, One year, Two year	Account	Predictor
PaperlessBilling	Categorical	Yes, No	Account	Predictor
PaymentMethod	Categorical	Electronic check, Mailed check, Bank transfer, Credit card	Account	Predictor
MonthlyCharges	Numeric	18.25 – 118.75 (USD)	Financial	Predictor
TotalCharges	Numeric	18.80 – 8684.80 (USD)	Financial	Predictor
Churn	Binary Categorical	Yes (1), No (0)	—	Target Variable

Source: IBM Telco Customer Churn Dataset, Kaggle

Annexure B: Churn Risk Segment Classification Rules

The following classification rules were derived from the exploratory data analysis findings documented in Chapter 4. These rules translate the empirical patterns identified in the data into operational customer segmentation logic that can be directly implemented in a CRM system using standard SQL or rule-engine logic.

Table A.2: Churn Risk Segment Definitions and Characteristics

Segment	Classification Rule	Avg. Churn Rate	Strategic Priority	Recommended Action
Critical Risk	Contract = Month-to-Month AND Tenure $\leq$ 12 months AND InternetService = Fiber optic	~55–60%	Highest	Immediate personal outreach; offer contract upgrade incentive and value-added service bundle
High Risk	Contract = Month-to-Month AND Tenure $\leq$ 12 months AND InternetService $\neq$ Fiber optic	~40–45%	High	Automated retention sequence; second-purchase incentive; online security trial offer
Moderate Risk	Contract = Month-to-Month AND Tenure > 12 months	~30–35%	Moderate	Loyalty recognition program; annual contract conversion offer with price incentive
Structural Risk	Contract = One Year AND Tenure $\leq$ 24 months	~11–15%	Moderate-Low	Pre-renewal outreach at month 10–11; two-year upgrade offer
Low Risk	Contract = Two Year OR Tenure > 48 months	~3–8%	Low	Standard retention program; focus on upselling value-added services
Safe	Contract = Two Year AND Multiple value-added	~2–4%	Minimal	Loyalty rewards; referral program enrollment

Segment	Classification Rule	Avg. Churn Rate	Strategic Priority	Recommended Action
	services subscribed AND Tenure > 36 months			

Source: Own Analysis based on exploratory data analysis findings, Chapter 4

Annexure C: SMOTE Variant Technical Specifications

The following table summarizes the technical specifications of the three SMOTE variants applied in this study, including the Python library used, key parameters, and the theoretical basis for synthetic sample generation.

Table A.3: SMOTE Variant Technical Specifications

Specification	Standard SMOTE	Borderline-SMOTE	SVM SMOTE
Python Library	imbalanced-learn	imbalanced-learn	imbalanced-learn
Class Reference	SMOTE()	BorderlineSMOTE()	SVM SMOTE()
k Nearest Neighbors	5 (default)	5 (default)	5 (default)
Sampling Strategy	Interpolation between all minority class instances	Interpolation between borderline minority instances only	Interpolation near SVM support vectors of minority class
Instance Selection	All minority class instances	Minority instances with $\geq 50\%$ majority class neighbors	Minority instances within SVM margin region
Synthetic Sample Placement	Random point on line segment between instance and random neighbor	Random point on line segment between borderline instance and random minority neighbor	Random point on line segment between support vector instance

Specification	Standard SMOTE	Borderline-SMOTE	SVMSMOTE
			and random minority neighbor
Theoretical Strength	Broad coverage of minority class feature space	Focused coverage near decision boundary	Globally optimal boundary characterization
Known Limitation	May generate noisy samples far from decision boundary	Heuristic boundary definition may not generalize	Higher computational cost due to SVM fitting
Target Sampling Ratio	1:1 minority to majority	1:1 minority to majority	1:1 minority to majority
Random State	Fixed (reproducibility)	Fixed (reproducibility)	Fixed (reproducibility)

Source: Chawla et al. (2002); Han et al. (2005); Nguyen et al. (200)

Annexure D: Complete Model Performance Results Matrix

The following table provides the complete performance results for all twelve model-resampling configurations evaluated in this study, serving as a single consolidated reference for all numerical findings reported across Chapter 4.

Table A.7: Complete Performance Results — All Model-Resampling Configurations

Resampling Strategy	Classifier	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Baseline (No SMOTE)	Logistic Regression	0.801	0.647	0.550	0.595	0.846
Baseline (No SMOTE)	Decision Tree	0.788	0.618	0.529	0.570	0.825

Resampling Strategy	Classifier	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Baseline (No SMOTE)	XGBoost	0.801	0.669	0.497	0.570	0.846
Standard SMOTE	Logistic Regression	0.755	0.526	0.773	0.626	0.843
Standard SMOTE	Decision Tree	0.730	0.495	0.818	0.617	0.816
Standard SMOTE	XGBoost	0.771	0.552	0.728	0.628	0.843
Borderline-SMOTE	Logistic Regression	0.753	0.523	0.769	0.623	0.842
Borderline-SMOTE	Decision Tree	0.743	0.511	0.717	0.597	0.809
Borderline-SMOTE	XGBoost	0.778	0.565	0.707	0.628	0.845
SVMSMOTE	Logistic Regression	0.764	0.539	0.758	0.630	0.845
SVMSMOTE	Decision Tree	0.742	0.509	0.747	0.606	0.813
SVMSMOTE	XGBoost	0.769	0.547	0.758	0.636	0.845

Source: Own Analysis

Annexure E: List of Abbreviations

Table A.9: List of Abbreviations Used in the Study

Abbreviation	Full Form
AUC	Area Under the Curve
CRM	Customer Relationship Management
CSV	Comma Separated Values
DTU	Delhi Technological University
DSL	Digital Subscriber Line

Abbreviation	Full Form
F1	F1-Score (Harmonic Mean of Precision and Recall)
FN	False Negative
FP	False Positive
GDPR	General Data Protection Regulation
IBM	International Business Machines Corporation
KPI	Key Performance Indicator
LIME	Local Interpretable Model-Agnostic Explanations
LR	Logistic Regression
MAE	Mean Absolute Error
MBA	Master of Business Administration
ML	Machine Learning
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
SVMSMOTE	Support Vector Machine Synthetic Minority Oversampling Technique
TN	True Negative
TP	True Positive
USD	United States Dollar
XGBoost	Extreme Gradient Boosting

Source: Own Compilation