Major Research Project On

“ENHANCING TELECOM RETENTION: LEVERAGING DATA SCIENCE TO MITIGATE CUSTOMER CHURN”

Submitted By:

Aman Singh Kushwaha

23/DMBA/15

Submitted to:

Dr. Chandan Sharma

Assistant Professor

DELHI SCHOOL OF MANAGEMENT

Delhi Technological University

Bawana Road Delhi 110042

Certificate

This is to certify that the project entitled " **Enhancing Telecom Retention: Leveraging Data Science to Mitigate Customer Churn** " has been completed by Aman Singh Kushwaha, 23/DMBA/15 to Delhi School of Management, Delhi Technological University, in partial fulfilment of the requirement for the award of the degree of Masters in Business Administration during the academic year 2024–2025.

Submitted to:

(Dr. Chandan Sharma)

Place:

Date:

Declaration

I, **Aman Singh Kushwaha**, student of MBA 2023-25 of Delhi School of Management, Delhi Technological University, Bawana Road, Delhi - 42, hereby declare that the research report " Enhancing Telecom Retention: Leveraging Data Science to Mitigate Customer Churn " submitted in partial fulfilment of Degree of Masters of Business Administration is the original work conducted by me.

The information and data given in the report is authentic to the best of my knowledge.

This report is not being submitted to any other University, for award of any other Degree, Diploma or fellowship.

Place: (Aman Singh Kushwaha)

Date:

Acknowledgement

I, **Aman Singh Kushwaha**, wish to extend my gratitude to Dr. Chandan Sharma, Asst. Professor Delhi School of Management (DSM), Delhi Technological University; for giving me all the direction and valuable insights to take up this Semester Project.

I also take this opportunity to convey sincere thanks to all the faculty members for directing and advising during the course.

I am equally grateful to my peers, friends, and family for their unwavering support, motivation, and understanding during the entire journey of this project.

This work would not have been possible without the contributions of each of the aforementioned individuals, and I express my deepest appreciation to all.

Executive Summary

With the rapidly changing telecom industry, retaining customers has become all the more significant due to rising competition, saturation in the market, and the ability of customers to switch service providers easily. Though customer acquisition costs continue to escalate, telecos now look towards retaining current customers so that the company remains profitable in the long run as well as sustainable. This project, "Enhancing Telecom Retention: Using Data Science to Fight Customer Churn," aims to solve the issue of customer churn by using advanced machine learning techniques in an attempt to spot trends in user behavior and predict likely churners.

The primary goal of this research is to develop a correct and trustworthy churn prediction model based on historical customer data. The research compares various machine learning algorithms like logistic regression, decision tree and random forest, in order to find the best approach for churn identification. The models are ranked in terms of performance metrics of accuracy, precision, recall, and F1-score.

The results of the study show that ensemble models, in this instance, Random Forest, are more predictive in nature than single classifiers. Furthermore, application of data balancing methods like SMOTE (Synthetic Minority Over-sampling Technique) drastically enhanced recall rates of the models to support better identification of at-risk customers. This suggests more accurate and cost-saving efforts at retention by preventing pointless efforts on low-risk clients.

The initiative pertains to the point that although churn prediction is useful, its real value stems from being a component of complete customer relationship management (CRM) initiatives. Even a minor increase in retention can lead to appreciable cost savings, in light of the high lifetime value of telecommunication customers.

This research not only illustrates the empirical application of machine learning to address real-world business issues but also delineates a model for organizations seeking to develop their churn management capacity. The findings validate the potency of data-driven choice making in producing competitive strength and inducing long-term customer engagement.

Chapter 1: INTRODUCTION

Churn Rate:

Churn rate, also known as attrition rate, is a business metric that measures the proportion of customers who stop doing business with a company over a specific time period. It is expressed as a percentage and is calculated using the formula:

$$\text{Churn Rate} = \left(\frac{\text{Customers Lost During a Period}}{\text{Total Customers at the Start of the Period}} \right) \times 100$$

For example, if a telecom company had 10,000 customers at the start of the month and lost 500 by the end, the churn rate would be:

$$\left(\frac{500}{10,000} \right) \times 100 = 5\%$$

Types of Churn

Understanding the nuances of churn is important for tailoring industry-specific strategies.

Common types include:

1. **Voluntary Churn**: Customers leave intentionally due to dissatisfaction, better alternatives, or pricing concerns.
2. **Involuntary Churn**: Caused by reasons beyond customer choice, such as failed payments or account closures due to technical errors.
3. **Customer Churn**: Specifically refers to the loss of end-users or clients.
4. **Revenue Churn**: Represents the loss in monetary value, which may differ even if customer numbers remain stable (e.g., when high-paying clients leave).

Why Churn Matters: Impact Across Industries

1. **Telecommunications Industry**
    - High churn directly reduces Average Revenue Per User (ARPU) and overall profitability.
    - Telecom markets are extremely saturated and competitive, and customer retention is therefore more desirable compared to acquisition.
    - Acquisition Cost vs. Retention Cost: 5–7 times the cost to retain a customer may be needed to acquire a new telecom customer.

    High churn could reflect poor quality of service, coverage problems, or inefficient customer support.

    Impact: Revenue loss, reduced brand loyalty, higher operational costs, and eroded market share.

2. **Software as a Service (SaaS)**
    - SaaS companies are highly dependent on subscription-based recurring revenues (yearly/monthly payments).
    - SaaS churn is especially costly since the Customer Lifetime Value (CLTV) is front-end biased; losing a customer early on eliminates future cash flows.
    - Positive churn (revenue from existing customers via cross-sells/upsells) is a big counterbalance.

    Impact: Slowed growth, diminished investor confidence, difficulty achieving profitability, especially in early-stage startups.

3. **Banking and Financial Services**
    - High churn indicates low customer engagement or dissatisfaction with service offerings.
    - Digital banking makes switching easier, increasing attrition risk.
    - High-net-worth individuals (HNIs) leaving has a disproportionate revenue impact.

    Impact: Loss of deposits, decline in asset under management (AUM), reputational damage, and reduced cross-selling opportunities.

4. **E-Commerce and Retail**
    - For online platforms, churn may appear as customer inactivity (not purchasing over a set period).
    - Retention drives repeat purchases, which are more profitable than one-time buys.
    - Personalized marketing and loyalty programs are crucial to mitigate churn.

    Impact: Reduced average order value, lower conversion rates, and greater customer acquisition pressure.

5. **OTT & Streaming Services**
    - OTT platforms like Netflix, Spotify, etc., are subscription-driven. Content quality, pricing, and user experience heavily influence churn.
    - Seasonal churn is common due to temporary subscriptions.

    Impact: Decline in Monthly Recurring Revenue (MRR), poor platform engagement, and difficulty in predicting future revenue.

6. **Insurance Industry**
    - Customer churn here affects long-term profitability because policies are usually renewed annually.
    - Higher churn can be due to pricing, claim processing inefficiency, or competitor offerings.

    Impact: Loss of renewal premiums, weakened underwriting pool, and lower brand trust.

7. **Hospitality and Travel**
    - Frequent churn indicates poor service quality or pricing mismatch.
    - With reviews and social media influence, churn can amplify reputational damage.
    - Loyalty programs are widely used to retain customers.

    Impact: Loss of future bookings, diminished online reputation, and loss of customer lifetime value.

General Business Implications of High Churn

1. **Lower Profitability:** Increased marketing spend to replace lost customers.
2. **Forecasting Challenges:** Inconsistent revenue streams hamper financial planning.
3. **Brand Reputation**: Word-of-mouth churn can discourage potential customers.
4. **Investor Relations**: For public or funded companies, high churn negatively affects valuation metrics like CLTV/CAC ratios.
5. **Employee Morale**: High churn often reflects systemic issues, creating internal friction.

Churn Rate as a Strategic Metric

Organizations track churn not only to plug leaks but to:

- Identify product/service deficiencies.
- Refine customer personas and preferences.
- Benchmark against competitors.
- Build personalized retention strategies.
- Optimize onboarding and customer service operations.

Decision Cycle of a Subscriber

The decision cycle is divided into two broad paths:

1. **Subscribers Who Have Not Considered Churning:**
    - **Inert Subscriber**
      These users continue with the service not out of loyalty, but due to inertia. They may find switching too time-consuming, complex, or not worth the effort.
    - **Unconditionally Loyal Subscriber**

      These users genuinely believe that their operator offers the best value or service, and they stay by choice, not by necessity.
    - **Locked-In Subscriber**

      These customers may want to switch but are bound by contractual obligations, such as penalties for early exit.

These three groups are generally stable, but Inert and Locked-In users may churn once constraints are lifted, making them important to monitor.

2. **Subscribers Who Have Thought About Churning:**
   These are more volatile and fall into four churn-related segments, each reflecting a deeper need for strategic intervention.
   - **Conditionally Loyal**
     They remain with the provider only because current conditions suit them. Their loyalty is fragile and dependent on ongoing offers or minor service satisfaction.
   - **Conditional Churner**

     They've decided to leave based on a better competing offer. Pricing and promotional strategies heavily influence this group.
   - **Lifestyle Migrator**

     These users churn due to changes in personal needs or life circumstances, not because of dissatisfaction. For example, moving to a new region with different service coverage.
   - **Unsatisfied Churner**

     This segment leaves due to persistent dissatisfaction—poor customer service, low network quality, or unresolved issues. They are the most critical to track and address using predictive analytics

Machine Learning

Machine learning is a branch of artificial intelligence that enables computers to learn from data and improve their performance on specific tasks without being explicitly programmed for each scenario. At its core, machine learning involves building models that can recognize patterns and make predictions or decisions based on input data.

The process begins with collecting and preparing data, which can include numbers, text, images, or other types of information. This data is used to train a machine learning model, allowing it to identify relationships and patterns. For example, if you provide a model with many examples of fruits-such as oranges, grapefruits, apples, and bananas-the model can learn to distinguish between them by analysing features like size, color, and shape.

There are several main types of machine learning:

- **Supervised learning:** The model is trained on labeled data, meaning each example in the training set is paired with the correct answer. The goal is for the model to learn to predict the correct label for new, unseen data. Common tasks include classification (assigning categories) and regression (predicting continuous values).
- **Ensemble learning:** it is a powerful machine learning approach that combines the predictions of multiple diverse models-such as decision trees, logistic regression, or neural networks-using techniques like bagging, boosting, or stacking, in order to achieve higher accuracy, greater robustness, and improved generalization compared to relying on a single model.

Random Forest Classifier

The Random Forest Classifier is a highly used machine learning algorithm for being applied to numerous classification issues like customer churn prediction in telecommunication. It is an ensemble algorithm in which an array of decision trees along with their output combination is created during the training and used to make a prediction. Output combination can be carried out by majority vote in classification problems, when the voted class most gets output. Random Forest exploits randomness in two fundamental manners: first, it applies bootstrap sampling to train every decision tree on an independent set of data; second, it chooses a random set of features at each split so that the trees are different. This blend of randomness and ensemble learning guards against overfitting and enhances the model's capacity to generalize. The algorithm has great precision, is noise robust, and can deal with missing values. For telecom churn prediction, Random Forest can identify subtle patterns of customer behavior, identify the most important drivers of churn, and offer meaningful observations on what drives customer decisions. Although it is more interpretable than other easier models and more computationally expensive, its ability to provide strong predictions makes it a superior starting choice for retention efforts that are data-driven.

Decision Tree Classifier

Decision Tree Classifier is a supervised learning algorithm that can be employed to address machine classification problems like customer churn prediction in the telecommunication sector. It partitions the data set into subsets and builds a tree structure in which a feature is an internal node of the tree, a decision rule a branch, and a class label a leaf node. The algorithm begins at the root node and successively applies decision rules to descend down the tree until it makes a prediction. To figure out what attribute to split on, measures such as Gini impurity or entropy (utilized in information gain) are utilized, which quantify the homogeneity of the target variable in the subsets. One of the biggest strengths of decision trees is that they are extremely easy to interpret and simple; rules produced are extremely interpretable and can be easily explained to stakeholders. Decision trees also have the capability of handling categorical and numerical variables, and need little data preprocessing. Decision trees, in churn prediction, also assist in determining the most significant attributes leading to customer churn, such as low usage, billing-related, or late complaint resolution issues. However, decision trees do overfit if they get too deep, but that can be avoided by techniques such as pruning. Otherwise, they form a good foundation for more advanced models such as Random Forests.

Imbalanced Learning

Imbalanced learning is a condition in machine learning when observations across classes are not evenly balanced. It occurs frequently in real-world scenarios such as customer churn prediction for telecommunication companies, where customers retained by the provider out-number customers who change service providers. In these cases, the default machine learning classifiers are biased toward the majority class, and the performance for minority class classification is poor, which is the class of greater significance in most scenarios. Suppose a model is efficient enough to classify all the customers to remain, and hence the overall accuracy will be greater, but it will be neglecting the actual churners. In order to overcome this skewness, there are some methods that can be employed. These involve resampling techniques such as oversampling the minority class (e.g., with SMOTE – Synthetic Minority Over-sampling Technique) or undersampling the majority class. Algorithm-level solutions like adjusting class weights or using dedicated algorithms for imbalanced learning become applicable in this case. Precision, recall, F1-score, and area under the ROC curve are superior measures of assessment compared to accuracy in this case. By unbalanced learning techniques, models are able to identify less common but crucial events more accurately, allowing telecom operators to forecast churn threats.

Evaluation Score

Machine learning models may not always be ideal for the provided data and need evaluation to see how well the model performs. Most standard ways of evaluating binary classifiers are implementing some steps which involve accuracy, recall, precision and F1-score. These metrics can be calculated through a confusion matrix

	Actual	
Predicted	True positive (TP)	False negative (FN)
	False positive (FP)	True negative (TN)

Confusion matrix is a machine learning term containing information about the actual and predicted classes, employed to depict the classifier's performance. True True Positives (TP) and True Negatives (TN) are the test samples properly classified while False Negatives (FN) and False Positives (FP) are the test cases that are misclassified.

Accuracy is a measurement that indicates the overall performance of the classifier. It is a measure indicating the ratio of total instances correctly classified. Accuracy, as reported by Deng et al is defined as:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

Accuracy is a measure showing the proportion of correctly predicted instances that were positive. The metric measures how frequently the model is accurate in predicting the target class, in our case churners. According to Deng et al. precision refers to the accuracy of predicting a specific class and calculated by the following:

$$\text{Precision} = \frac{TP}{TP+ FP}$$

Recall is a metric, which indicates the extent to which the classifier performs in locating examples denoted as positive. It represents the ability of the binary classifier in identifying instances of a specific class. Recall is computed as:

$$\text{Recall} = \frac{TP}{TP+ FN}$$

The F-score may be employed for the assessment of the accuracy of a classifier. F-score is a measure that considers both precision and recall and is generally characterized by the harmonic mean of recall and precision. Both recall and precision are improved as The. F-score is nearer to 1.

$$\text{F-score} = \frac{2 * precision * recall}{precision+recall}$$

Chapter 2: LITERARTURE REVIEW

1. Traditional Machine Learning Approaches

Early studies focused on supervised learning techniques like logistic regression, decision trees, and random forests for churn prediction. For instance:

- Kiran Dahiya and Surbhi Bhati (2015) demonstrated decision trees and logistic regression achieve 70-80% accuracy in identifying at-risk customers by analysing usage patterns and contract details.
- Recent comparisons show XGBoost and Random Forest outperform K-Nearest Neighbours and Support Vector Machines, achieving >80% accuracy and F1-scores in telecom datasets. These models excel at handling structured data like billing history and service usage.

2. Addressing Class Imbalance

Churn datasets are inherently imbalanced (e.g., 5% churn rates). Key strategies include:

- Oversampling techniques (e.g., SMOTE) to balance minority classes.
- Cost-sensitive learning to penalize misclassifications of churners more heavily.
- Hybrid models like BiLSTM-CNN (Bidirectional LSTM + Convolutional Neural Networks) improve recall by capturing temporal patterns in customer behaviour.

3. Deep Learning and Hybrid Models

Advanced architectures address limitations of traditional ML:

- BiLSTM-CNN (proposed by Saha et al., 2023) achieves 99% accuracy by combining sequential data analysis with feature extraction.
- Stacking ensemble models (e.g., XGBoost + logistic regression) improve robustness, achieving 81.05% accuracy by leveraging meta-learners to refine predictions.

4. Feature Engineering and Segmentation

- RFM (Recency, Frequency, Monetary) analysis simplifies dynamic features (e.g., call duration, data usage) to improve model interpretability.
- Micro-segmentation (clustering customers into 50+ groups) enables personalized retention strategies, reducing churn by 10–15%.
- Studies highlight contract type, tenure length, and monthly charges as top churn predictors Studies highlight contract type, tenure length, and monthly charges as top churn predictors.

5. Ethical and Operational Challenges

- Data quality issues: Missing values and inconsistent data collection hinder model performance.
- Privacy concerns: Balancing customer data utilization with GDPR compliance remains critical.
- Dynamic consumer trends: Models require continuous retraining to adapt to shifting market conditions.

6. Industry Applications

- Telecom giants like Jio and AT&T use real-time dashboards to visualize churn risk scores, enabling proactive interventions
- Case studies show feature discovery (identifying 50+ churn drivers like device-service mismatches) boosts retention efficiency by 20%

Chapter 3: RESEARCH DESIGN AND METHODOLOGY

PROBLEM STATEMENT:

Customer churn is a business-critical issue in the telecommunications sector with a tangible influence on revenue, operational performance, as well as customer lifetime value. In an over-saturated and highly competitive market, telecom service providers find it challenging to retain customers, particularly when confronted with short tenure, unsuitable pricing models, as well as unavailable service bundling. This initiative seeks to utilize data science methods to determine the most powerful causal causes of customer churn, create predictive models, and provide actionable retention recommendations that allow telecommunication businesses to act pre-emptively and reach out to vulnerable customers and curb churn rates.

OBJECTIVES OF THE RESEARCH:

1. To examine customer demographic, behavior, and service usage data to find important patterns and trends regarding customer churn for the telecommunication industry.
2. To identify the most important drivers of customer churn, such as contract types, payment methods, tenure, monthly charges, and value-added services.
3. To develop a predictive model based on data science tools (e.g., logistic regression, decision trees, or random forest) that effectively classifies customers as churned or retained.
4. To stratify customers by their probability of churn and profile each stratification to understand their unique characteristics and requirement for retention.
5. To provide actionable advice and data-driven strategies for telecommunications operators to reduce churn systematically and improve customer retention.

PROBLEMS IDENTIFIED:

1. Churn-related datasets are usually imbalanced, with a smaller proportion of churned customers, making traditional modelling less effective without proper preprocessing techniques.

2. There is a lack of detailed segmentation of customers based on churn risk, leading to inefficient or irrelevant retention efforts.
3. EDA often reveals churn correlations, but businesses fail to translate these into practical steps for customer engagement and loyalty programs.

RESEARCH DESIGN:

This research follows an applied and quantitative research strategy to tackle customer churn in the telecom industry through data science methods. The research utilizes formalized customer data such as demographics, service usage, and payments made. Data are preprocessed, cleaned, and reformed through exploratory data analysis (EDA) to identify patterns and major drivers of churn like contract type, tenure, payment, and monthly rate.

Supervised machine learning algorithms like Logistic Regression, Decision Trees and Random Forest, are trained to predict churn. Accuracy, recall, precision, and ROC-AUC model performance metrics are used to evaluate models. Model interpretability tools like SHAP and feature importance are used to interpret model predictions.

Analysis culminates in customer segmentation and actionable suggestions to improve retention—like long-term offers or bundling of services. Technology used is Python (Pandas, Scikit-learn, Seaborn) on Jupyter Notebook. The output will be capable of assisting telecommunication companies to proactively manage churn and promote customer loyalty.

METHODOLOGY:

The study is quantitative as it utilizes machine learning to forecast and minimize telecom customer churn. The data contains customer details, usage of services, type of contracts, billing details, and whether they have churned. Data cleaning to deal with missing values and inconsistencies is the initial step. Categorical features' encoding and scaling of numeric attributes are the next steps. To deal with class imbalance, methods like SMOTE are used.

Exploratory Data Analysis (EDA) is carried out to identify the patterns and trends between churn and significant variables like tenure, monthly charge, and type of contract. Different machine learning models—Logistic Regression, Decision Tree are

Random Forest, are cross-validated and trained. Accuracy, precision, recall, F1-score, and ROC-AUC are the performance metrics.

Feature importance methods make it possible to explain model predictions and determine key churn drivers. Ultimately, customer segments are constructed based on what has been discovered and actionable recommendations like contract rescheduling, service bundling, and targeted promotions are suggested to minimize churn and maximize retention.

LIMITATIONS OF THE STUDY-

1. The research is based on sample or publicly available data that does not necessarily cover all actual telecom variables or current market trends.
2. Predictor variables such as key qualitative and behavioural variables (e.g., complaint history, social sentiment, customer satisfaction) are not recorded, potentially lowering the performance of models.
3. The model depends on a point-in-time picture of the data; churn behaviour can be dynamic and subject to seasonality or time-varying effects.
4. The model presumes customers make decisions on the basis of quantifiable attributes, disregarding emotional or illogical churn drivers.

Chapter 4: DATA ANALYSIS AND INTERPRETATION

Data Collection and preparation

The dataset for this research was obtained from Kaggle, and it is specifically preprocessed for telecom customer churn data analysis. The dataset has 7,043 customer records, and each record corresponds to one telecom user. The data comprises demographic information, account information, and service usage patterns, and thus it is very appropriate for predictive modeling and customer behavior analysis.

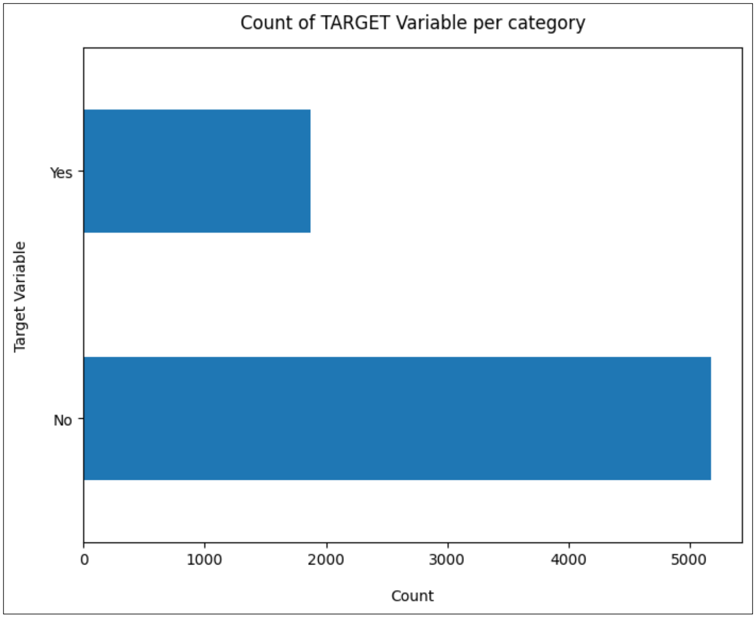
One of the more substantial variables in the data is 'Churn', which is a binary variable and represents whether or not the customer has churned (dropped the service). Of 7,043 rows:

5,174 customers (about 73.5%) have not churned; i.e., they are still on the telecom service.

1,869 customers (about 26.5%) have churned; i.e., they have dropped their service.

This split demonstrates a moderate class imbalance, which is critical in machine learning modeling. The sample provides a good baseline for characterizing customer churn patterns and causes, and developing retention strategies through data science methods.

Figure 4.1



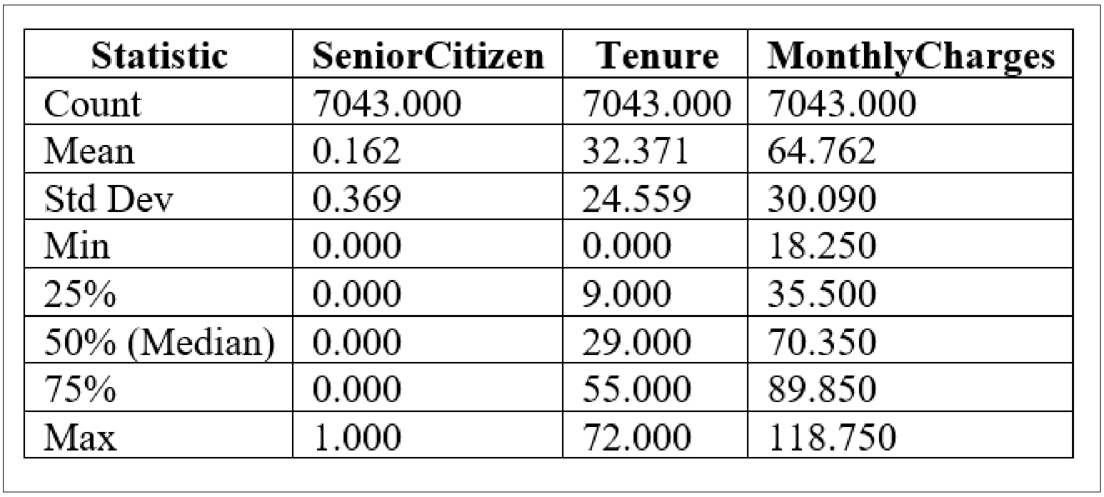
Source: Own Analysis

The data set contains 7,043 telecom customers. 'SeniorCitizen' is a binary variable indicating that 0 = non-senior citizens and 1 = senior citizens. The mean of 0.162 reveals that about 16.2% of the customers are senior citizens and the rest 83.8% are non-senior citizens. The 25th, 50th, and 75th percentiles all equalling 0 also ensures that most customers are non-senior citizens.

The 'tenure' variable shows the length of time the customer has been with the telecommunication organization. The mean tenure is 32.37 months, ranging from a minimum of 0 to a maximum of 72 months. A 24.56-month average standard deviation implies the existence of a high range of customer duration. Most notably, 25% of customers have tenure of less than 9 months, 50% have tenure less than 29 months, and 75% have tenure less than 55 months, meaning that the majority of customers are fairly recent.

The 'MonthlyCharges' variable is ₹64.76 on an average, ranging from ₹18.25 to ₹118.75. Its standard deviation is 30.09, reflecting high volatility in billing. The 25th percentile is ₹35.50, the median is ₹70.35, and the 75th percentile is ₹89.85. This reflects that the majority of customers have a moderate to high monthly charge with a significant number of customers over ₹70 per month.

Table 4.1

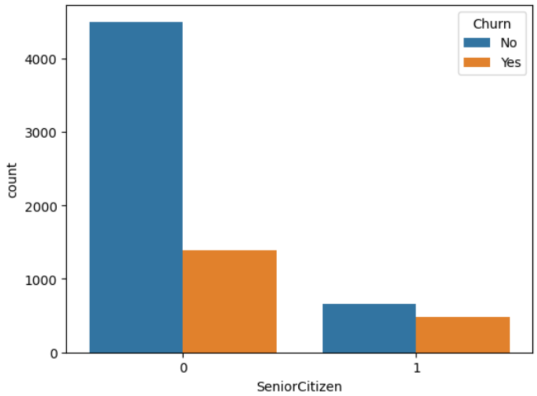


Statistic	SeniorCitizen	Tenure	MonthlyCharges
Count	7043.000	7043.000	7043.000
Mean	0.162	32.371	64.762
Std Dev	0.369	24.559	30.090
Min	0.000	0.000	18.250
25%	0.000	9.000	35.500
50% (Median)	0.000	29.000	70.350
75%	0.000	55.000	89.850
Max	1.000	72.000	118.750

Source: Own Analysis

Univariate Analysis

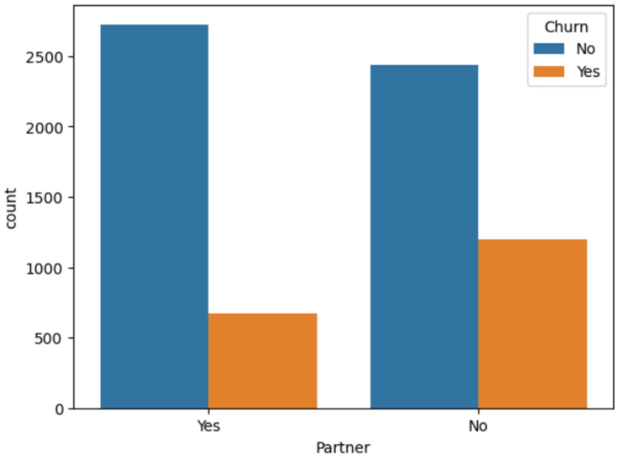
Figure 4.2



Source: Own Analysis

This graph illustrates that elder citizens (1) have a higher proportion of churning customers compared to non-elder citizens (0). Although most customers are not elderly citizens, the senior citizen churn rate is evident and high, which means that age is an important factor for customer loss in telecommunication services

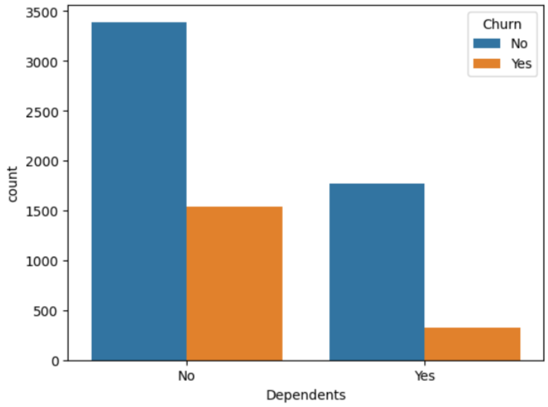
Figure 4.3



Source: Own Analysis

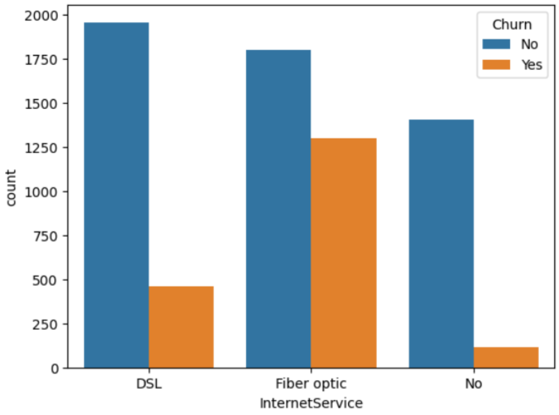
This graph illustrates how customers who don't have a partner are likely to churn more than their partners. The churned customer count is extremely greater for customers without a partner, indicating that partnership is related to higher customer retention in the telecom industry

Figure 4.4



Source: Own Analysis

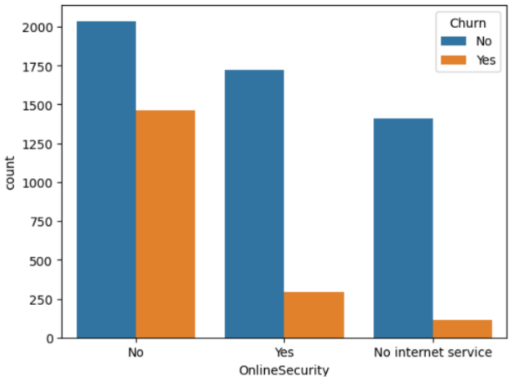
This graph shows that the customers with no dependents are more likely to churn than customers with dependents. Customers with no dependents have a much greater number of churn, which demonstrates that family obligations could be a determinant of loyalty and a reduced churn among telecommunication customers



Source: Own Analysis

The customers of fiber optic internet service are having a significantly higher number of churn than DSL or no internet service customers. Although DSL and fiber optic have the same numbers of customers who did not churn, the churned customers are higher in number for the fiber optic customers. The no internet service customers are having the lowest number of churn. This trend indicates that customers of fiber optic service are more likely to churn, perhaps owing to having a higher expectation level or price or service quality dissatisfaction.

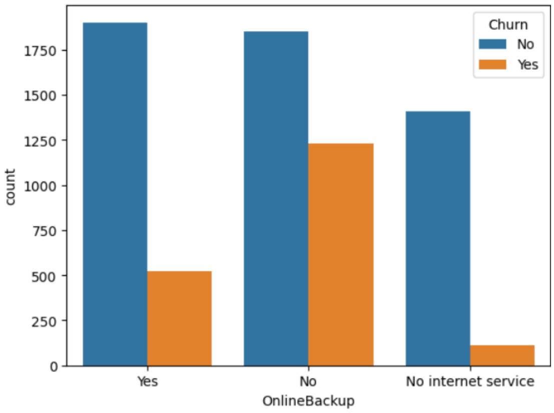
Figure 4.5



Source: Own Analysis

Customers who are not satisfied with online security are more likely to churn than their counterparts who own this facility. Churned customers are far more numerous in the group without online security, and among customers who have online security, much fewer have churn. Customers with no internet service at all have the minimum churn among these groups. This shows that the provision of online security features can minimize churn because customers whose data feels secure are more likely to remain.

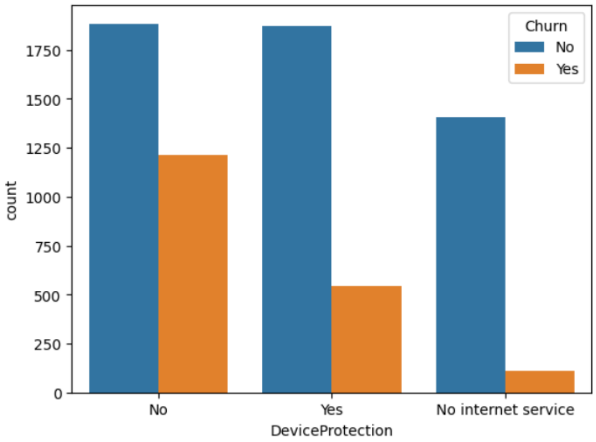
Figure 4.6



Source: Own Analysis

Non-online backup clients churn more than clients using it. There are higher churn figures among non-online backup clients, while clients who use the service have fewer churned clients. Similar to the previous trends, customers who lack access to the internet have the least churned clients. What this indicates is that providing online backup as a value-added service can enhance customer retention because customers like having extra features that safeguard their data.

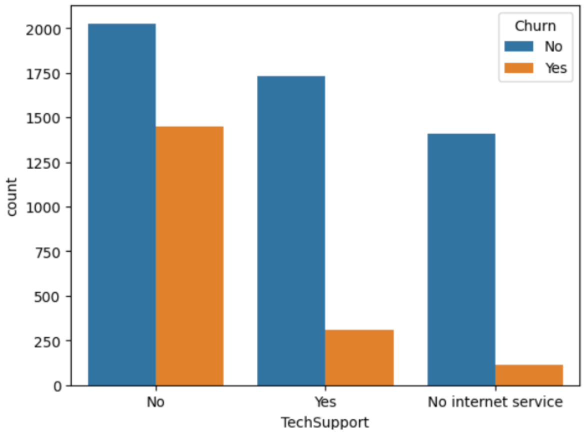
Figure 4.7



Source: Own Analysis

This graph indicates that customers without device protection churn higher than device protected customers. Out of the customers who do not have device protection, the value of churn customers is extremely high compared to customers who have this service. Customers who possess device protection possess a much lower value of churn. Customers who do not possess an internet service are in the lowest churn among them. This is to say that providing device protection will deflect churn as it is value add and customer protection.

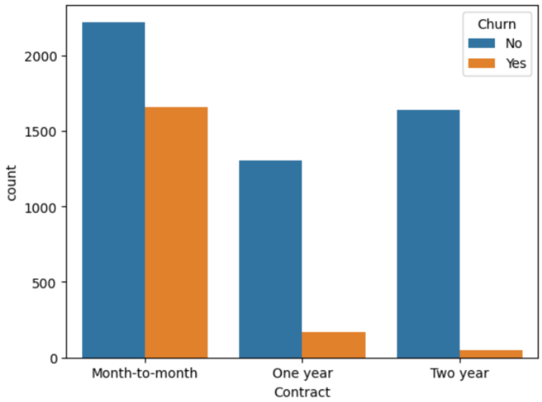
Figure 4.8



Source: Own Analysis

This chart illustrates how customers without tech support tend to churn more. The churn of customers without tech support is significantly higher compared to customers with tech support. Customers with tech support have significantly less churn, while those without internet service have the least churn again. This highlights the importance of providing good, accessible tech support as a customer retention method.

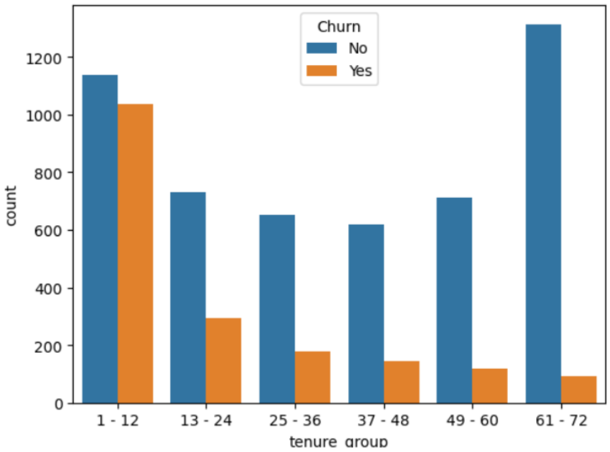
Figure 4.9



Source: Own Analysis

This graph shows a apparent correlation between contract type and churn percentages. Month-to-month customers have the highest churn, i.e., they are most likely to be canceled. One-year and particularly two-year customers have much lower churns. This trend indicates that longer commitment guarantees customer holding as well as lower churn.

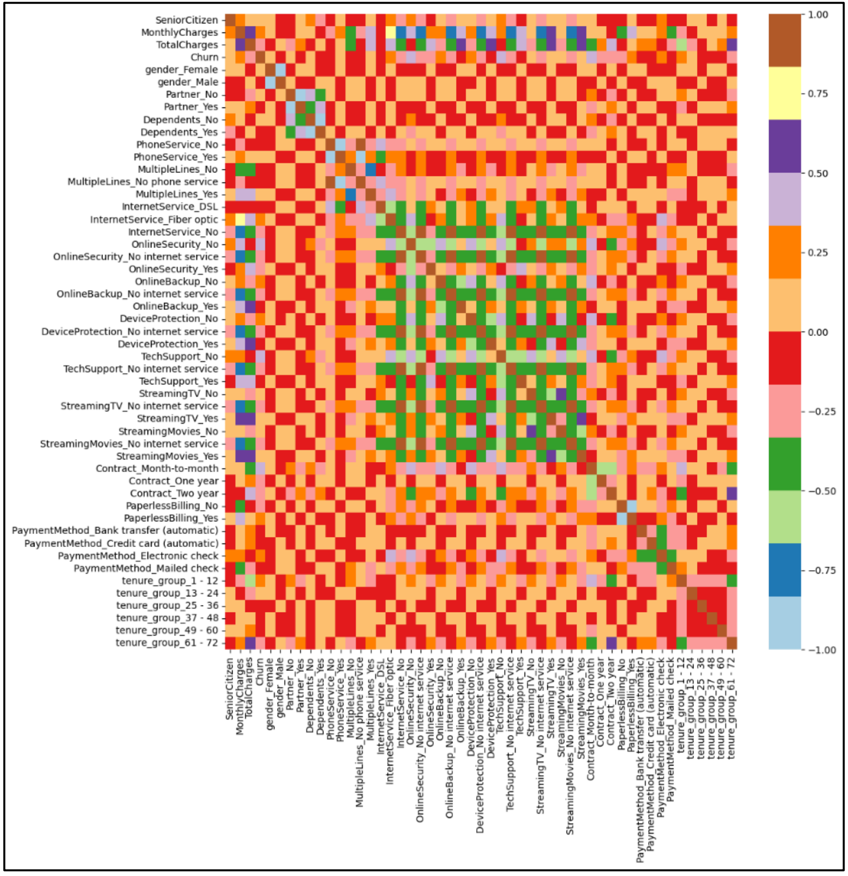
Figure 4.10



Source: Own Analysis

The fourth graph indicates that the largest churn rate exists among customers in the first tenure category (1–12 months) where churn numbers are roughly equivalent to retained customers. The numbers of churned customers decrease remarkably as tenure increases across all subsequent categories. Customers with the longest tenure (61–72 months) have the lowest churn rate. This implies that new customers are likely to switch, but long-term customers are most likely to stay, which indicates the importance of early intervention and retention activity

Figure 4.11



Source: Own Analysis

The heatmap presents two-year deals highly negatively correlated with churn, which verifies the customers with two-year deals are significantly less likely to churn away from the company. Month-to-month deals are positively correlated with churn, which tells us such customers are the most perilous segment to churn.

Monthly fees are positively correlated with churn, reflecting that higher fees are linked with greater customer defection. This price sensitivity is a critical consideration in retention strategy design.

Value-added features such as OnlineSecurity, OnlineBackup, and TechSupport are negatively correlated with churn, reflecting these features are utilized as tools of retention by enhancing customer stickiness.

Fiber internet service is positively related to churn even as a premium service, which implies that there may be service quality or price issues that need to be addressed.

The tenure-related variables are negatively related to churn with strong effect, which verifies that customer loyalty does grow strongly with duration and that initial relationship durations are most susceptible.

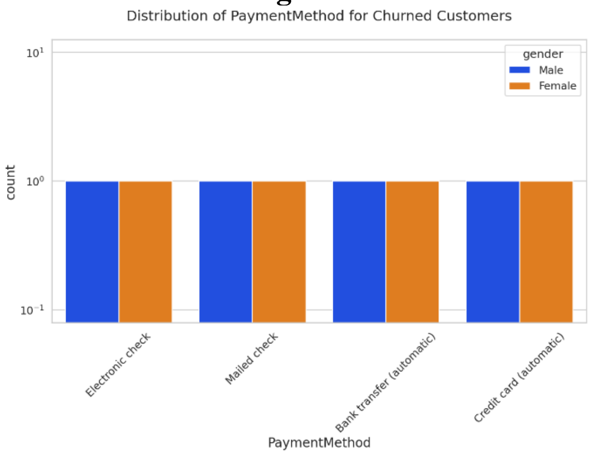
Demographic factors identify that Partner_No and Dependents_No are positively correlated with churn, i.e., customers who do not have family relationships tend to churn.

Internet service attributes display patterns of inter-correlated relationships, and TV and Movies streaming services provide high inter-correlations, indicating customers usually buy these services in a package.

Payment mechanism variables are less correlated with churn than contract type and service features, which means that payment mechanisms play a lesser role in contributing to driving retention outcomes compared to service quality and contractual commitment dimensions.

Bivariate Analysis

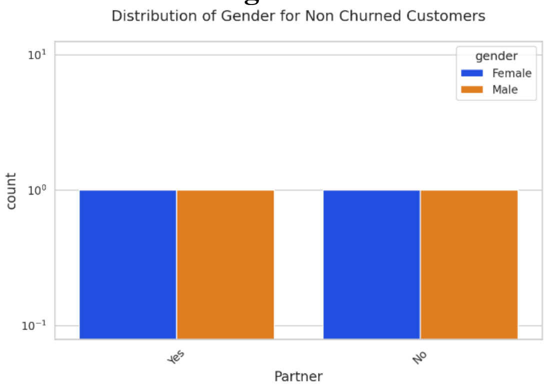
Figure 4.12



Source: Own Analysis

This graph shows the Distribution of PaymentMethod for Churned Customers and shows that churn rates are fairly stable across payment methods when comparing male to female customers. A log axis displays similar ratios of male and female customers churning in each payment category (electronic check, mailed check, bank transfer, and credit card). This indicates gender does not have a significant interaction with payment method to drive churn behavior, whereas the distribution in general indicates electronic checks could potentially have higher absolute churn counts, which is consistent with industry observations that electronic check customers are likely to experience higher churn rates (about 36.88%) than credit card customers (14.48%).

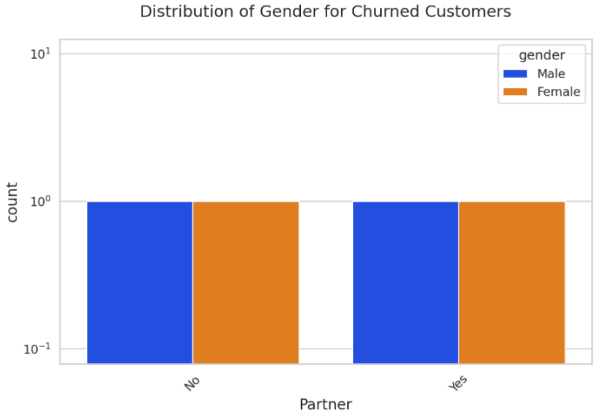
Figure 4.13



Source: Own Analysis

This figure illustrating Distribution of Gender for Non-Churned Customers illustrates that retention trends are the same among females and males independent of partner status. Logarithmic scale illustrates similar numbers of female and male customers that have stuck with the service provider, wherein both partner groups (Yes/No) exhibit similar gender distributions. This means that among customers who remain with the firm, gender and the presence of partners are non-discriminative distinction factors in churn behavior.

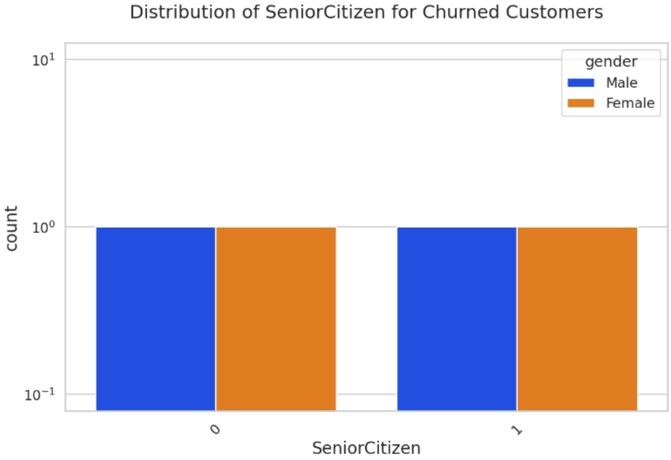
Figure 4.14



Source: Own Analysis

This chart that depicts Distribution of Gender for Churned Customers depicts the way churned customers who have left the firm have been divided by gender and partnership status. The image reveals that churn activity is uniformly distributed between male and female customers regardless of their partner status. This balanced division implies that gender has relatively little influence on churn decision even after controlling for relationship status. This is important because although partner status alone has been found to influence churn rates (customers without a partner are more likely to churn), its interaction with gender does not seem to be zero.

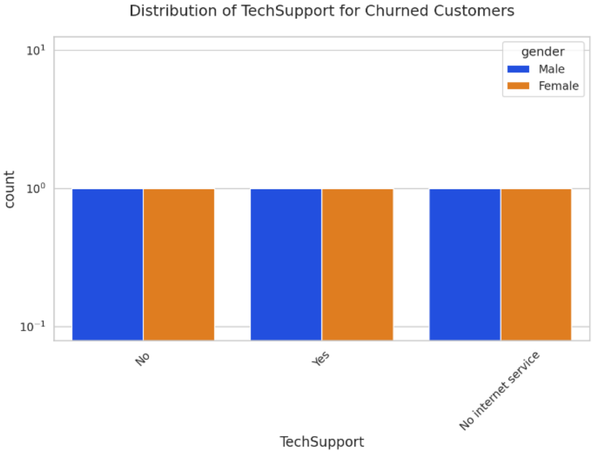
Figure 4.15



Source: Own Analysis

The SeniorCitizen Distribution for Churned Customers graph indicates that gender plays no role in influencing churn behavior among senior citizen segments and male and female customers exhibit the same churn both among senior citizen segments (0 and 1). Logarithmic scale has been adopted to establish equal sex ratios between the two genders, implying that senior citizen status remains a significant churn driver from earlier analyses but gender is not creating significant differences within such segments.

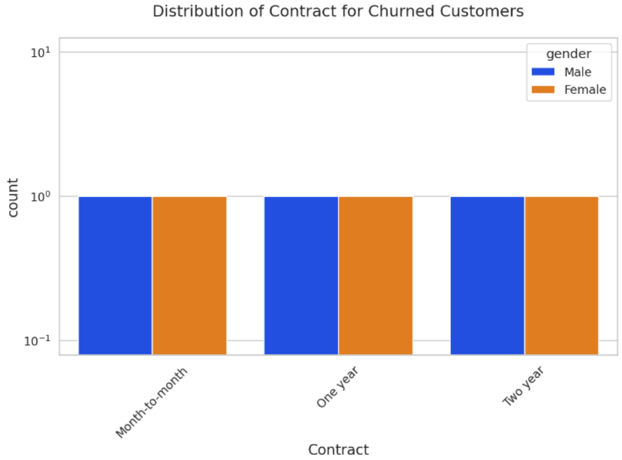
Figure 4.16



Source: Own Analysis

The Distribution of TechSupport for Churned Customers graph shows that technical support subscription status also affects churn the same way in males and females. Female and male customers reflect almost identical churn percentages for each technical support status (No, Yes, No internet service). This reflects that while tech support availability makes a difference in overall retention rates, gender is not a discriminator in how it makes a difference in churn, showing that value perceptions of the service cross over gender differences.

Figure 4.17



Source: Own Analysis

The Churned Customers Distribution of Contract chart shows contract type has an impact on churn behavior by gender similarly throughout. The balanced bars both being the same height for male and female on all contract types (Month-to-month, One year, Two year) supports gender does not significantly interact with contract type to decide the probability of churn. It is as previously evidenced, contract format itself is a primary cause of churn and its effect is one of the same use across gender segments.

Algorithm Choice

Since the type of problem—telecom customer churn prediction—is one of binary classification, the problem is not a regression one. Hence, regression-based solutions are disqualified. Algorithms like Logistic Regression, Naïve Bayes, Support Vector Machines, Decision Trees, and ensemble algorithms like Random Forest are commonly used in classification problems.

For the sake of this research, we have chosen the Decision Tree Classifier and Random Forest Classifier as they have been proven to work well in previous churn prediction studies and can cater to categorical and mixed-type data. The Decision Tree Classifier offers the strength of an understandable and interpretable model with which the impact of certain features on the chance of churn is easy to understand. In comparison, Random Forest Classifier, an ensemble learner that is based on bagging of high numbers of decision trees, gives enhanced precision and stability with less overfitting.

With the inclusion of a transparent and low-variance model (Decision Tree) and a powerful ensemble learner (Random Forest), the research combines transparency and efficacy to deliver accurate churn prediction alongside actionable information.

Parameter and Model Selection

Machine learning algorithms tend to necessitate careful hyperparameter tuning involving numbers of estimators, maximum depth, and splitting criterion to function perfectly. Hyperparameter tuning, although valuable, is both time-consuming and needs vast computational resources.

For the evaluation of this research, to leverage the machine learning algorithms in customer churn prediction in the telecommunication sector, we decided to use default parameters from the scikit-learn module for Decision Tree Classifier and Random Forest Classifier. The major focus of the research is to validate the effectiveness of these models in marking potential churners, not necessarily attaining optimal performance through high-scale parameter tuning.

Hence, optimization of parameters was left out as outside the purview of this research. We used default parameters in trying to establish the baseline performance of models selected and to prevent complexity and unwanted complicacies during the implementation process.

Training

Supervised machine learning training is a method that employs labeled data—i.e., input features and target output label—and is utilized to train a prediction model. For our telecom customer churn prediction task, we employed supervised learning since the dataset is labeled with customer features and their respective churn status (Churn/Not Churn).

For the assessment of the model performance, the data was split into a training set and a test set. It enables the model to learn from one set of data (training set) and test on another (test set) for generalizability. The division makes the model testing unbiased and fair.

Our dataset is unbalanced, with the majority being non-churners and the rest churners. We addressed this by resampling the training data. We used SMOTE for oversampling the minority class, RandomUnderSampler for undersampling the majority class, and SMOTEENN as an ensemble of both. The procedures ensured the data was balanced in a way that the model learns from both churned and non-churned classes equally, enhancing classification accuracy and minimizing bias against the majority class.

Evaluation

It is an important step of the machine learning process whereby it measures how well the model that was trained performs on new data. In this research where we are predicting telecom customer churn, we compared Decision Tree Classifier and Random Forest Classifier performance using common classification metrics—accuracy, precision, recall, and F1-score—derived from the confusion matrix.

Since the data set is unbalanced and has significantly more non-churners than churners, accuracy can no longer be a good estimator of model performance alone. So, precision and recall were given more importance, and most importantly, the F1-score since it is a harmonic mean of recall and precision.

Among those, recall was considered more precious in our case, since identifying actual churners is more important than reducing false positives. The test set was employed for evaluation, which the model never saw during training and hence gave an objective estimation of the generalization capacity of the model.

Implementation

We deployed our machine learning models in Python with a general collection of robust libraries for model building and data manipulation. Pandas and NumPy were used for data manipulation and preprocessing, while data visualization was facilitated by Matplotlib and Seaborn to identify trends and patterns. We utilized Scikit-learn, a robust Python library containing ample classification tools along with model testing and model building tools, to perform model testing and model building. The in-built functions of Scikit-learn enabled us to use algorithms like Decision Tree Classifier and Random Forest Classifier with ease and also perform data splitting, resampling, and calculate metrics. Utilization of all these libraries facilitated an efficient and reproducible process through the entire analysis and model-building process.

Chapter 5: FINDINGS AND CONCLUSION

Here, we introduce the performance result of the two chosen classifiers, Decision Tree Classifier and Random Forest Classifier. The models' overall performance is compared on the basis of the most important measures that include accuracy, F1-score, and correct classifications.

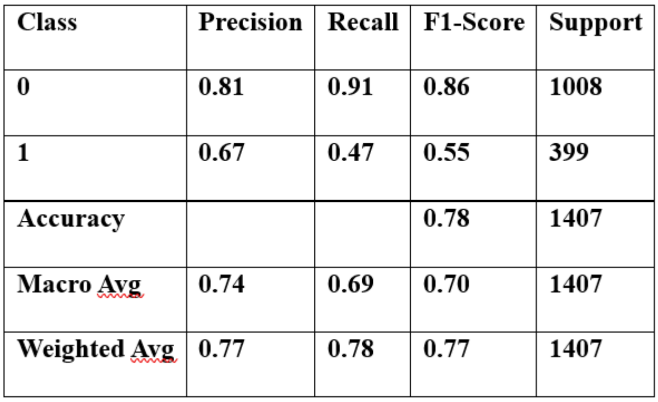
To understand how well the models predict the target class—Churn—we examine the confusion matrix and the corresponding metrics such as precision, recall, and F1-score. The confusion matrix is categorized into four parts:

- True Positives: Properly predicted churners,
- False Positives: Non-churners wrongly predicted as churners,
- False Negatives: Churners wrongly predicted as non-churners,
- True Negatives: Properly predicted non-churners.

All the classifiers were trained on the raw imbalanced dataset and on balanced datasets using sampling methods like SMOTE, RandomUnderSampler, and SMOTEENN to see the effect of balancing on the performance of the model.

Decision Tree Classifier

Table 5.1



Class	Precision	Recall	F1-Score	Support
0	0.81	0.91	0.86	1008
1	0.67	0.47	0.55	399
Accuracy			0.78	1407
Macro Avg	0.74	0.69	0.70	1407
Weighted Avg	0.77	0.78	0.77	1407

Source: Own Analysis

Decision Tree Classifier was likewise trained and tested using the same original imbalanced dataset, with most of the instances being from the non-churn class. The model produced a grand total of 78% accuracy because it was able to classify 78% of the entire 1407 test cases. Accuracy, however, is not enough to quantify because there is a class imbalance in the data set.

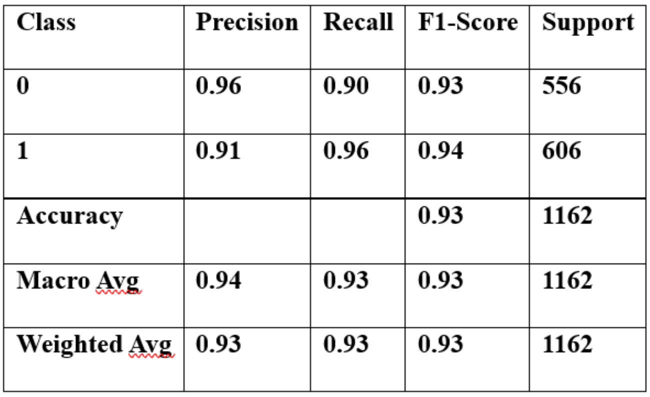
In class-wise performance, the model worked well for the non-churn class (class 0) with precision = 0.81, recall = 0.91, and F1-score = 0.86. This informs us that the model does well in correctly labeling non-churners and not creating false positives for this class. Its performance for the churn class (class 1) was quite poor, though. The accuracy of the churners was 0.67, that is, only 67% of the model-predicted churners actually churned. The recall was 0.47, which means that only 47% of the true churners were identified as such by the model, and the other 53% were missed. The F1-score for the churn class was 0.55, which is quite a low score and a measure of how badly the model can identify churn cases.

The macro average F1-score, which is class-balanced and gives equal weighting to both classes, was 0.70, while the weighted average F1-score, class support-aware, was 0.77. These also point out that the model performance is skewed in favor of the majority class.

Finally, Decision Tree Classifier had adequate results for non-churn class but underperformed for churn class. This variation in predictive capacity indicates that it would be crucial to use resampling methods like SMOTE so that the model would be able to accurately identify churners.

Applying SMOTEENN

Table 5.2



Class	Precision	Recall	F1-Score	Support
0	0.96	0.90	0.93	556
1	0.91	0.96	0.94	606
Accuracy			0.93	1162
Macro Avg	0.94	0.93	0.93	1162
Weighted Avg	0.93	0.93	0.93	1162

Source: Own analysis

After the SMOTEENN resampling technique had been applied to address class imbalance in the data, the accuracy of the Decision Tree Classifier improved significantly. The model was validated against a balanced data of 1162 instances, which included 556 non-churners (class 0) and 606 churners (class 1). It achieved an overall accuracy rate of 93%, indicating that 93% of the total instances were correctly classified.

For class 0 (the non-churn class), the model performed precision at 0.96, where 96% of the selected non-churners were accurate. The recall was at 0.90, where 90% of the actual non-churners were selected by the model. The F1-score for the class was at 0.93, which describes good trade-off between precision and recall.

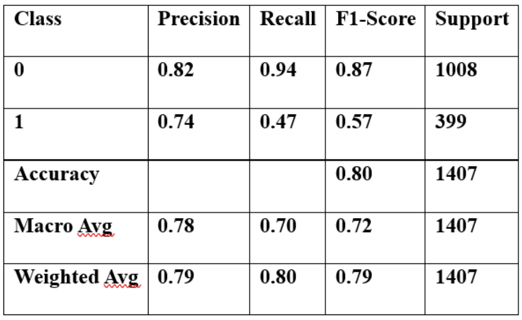
Likewise, for churn class (class 1), the model was performing great with precision 0.91 and recall 0.96. This indicates that 91% of customers which were predicted to churn were actually churned, and 96% churners were identified correctly. The F1-score of 0.94 in class indicates the model's capability to identify churn behavior.

The macro-average F1-score was 0.93, and it depicted balanced performance for both classes. The weighted average F1-score was also 0.93, which vindicated the impartiality of the model even with respect to class distribution.

Overall, the use of SMOTEENN significantly improved the classifier in identifying churners with extremely high accuracy in both classes. This validates that addressing class imbalance using advanced resampling techniques can have a tremendous effect on enhancing the predictive power of machine learning algorithms in churn prediction tasks.

Random Forest Classifier

Table 5.3



Class	Precision	Recall	F1-Score	Support
0	0.82	0.94	0.87	1008
1	0.74	0.47	0.57	399
Accuracy			0.80	1407
Macro Avg	0.78	0.70	0.72	1407
Weighted Avg	0.79	0.80	0.79	1407

Source: Own Analysis

Random Forest Classifier, when run on the class-imbalanced dataset without applying SMOTEENN, provided an accuracy of 80% i.e., 80% of all 1407 instances were predicted by the model as correctly belonging to their respective classes. The test dataset comprised 1008 non-churners (class 0) and 399 churners (class 1) and thus depicts a large class imbalance.

For class 0, the non-churn class, the model performed outstandingly well with precision being 0.82, indicating that 82% of the customers the model had labeled as non-churners were correct. The recall was 0.94, indicating that 94% of the true non-churners were captured correctly by the model. The F1-score for this class was 0.87, indicating complete balance between precision and recall.

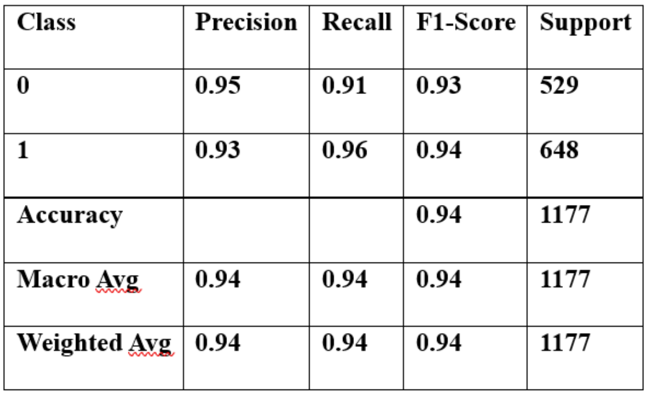
But the model did not do well on the churn class (class 1). It had a precision of 0.74, meaning 74% of the predicted churners were indeed churners. The recall decreased to 0.47, accounting for only 47% of the actual churners being correctly identified. The F1-score for class 1 was 0.57, indicating quite poor performance in identifying churners, with most of the weakness being explained by class imbalance.

Macro average F1-score was 0.72 and it measures the performance over the two classes without adjusting for class distribution. Weighted average F1-score was 0.79 and it attempts to balance the majority class by assigning higher weights.

Finally, despite the fact that the Random Forest Classifier was highly accurate as a non-churner, the detection of true churners was impaired by the imbalance nature of the data. The outcome of this makes the use of resampling methods like SMOTEENN which balance the data to enhance the model's capability in identifying churn behavior suitable.

Applying SMOTEENN

Table 5.1



Class	Precision	Recall	F1-Score	Support
0	0.95	0.91	0.93	529
1	0.93	0.96	0.94	648
Accuracy			0.94	1177
Macro Avg	0.94	0.94	0.94	1177
Weighted Avg	0.94	0.94	0.94	1177

Source: Own Analysis

After applying SMOTEENN resampling to the data, the Random Forest Classifier's performance significantly increased based on all the metrics used for evaluation. The overall accuracy of the model was around 93.80%, meaning that nearly 94% of all the 1177 instances were predicted correctly.

For class 0 (non-churn class), the model was precise to 0.95, i.e., 95% of those labeled as non-churners were correct. The recall was 0.91, showing that 91% of the true non-churners were correctly identified. The corresponding F1-score of 0.93 indicates a balance between recall and precision.

For class 1, the churn class, the classifier was equally good with precision at 0.93 and recall at 0.96. This indicates that 93% of the predicted churners were indeed churners and 96% of the true churners were identified. The F1-measure of 0.94 also establishes the strength of the model to identify churners after the class imbalance was corrected.

Macro average F1-score was 0.94, signifying that both classes performed similarly. The weighted average F1-score was 0.94 as well, which further signifies the model performing exceptionally even when class distribution is taken into account.

The use of SMOTEENN greatly enhanced the performance of the Random Forest Classifier in the accurate classification of churners and non-churners. The high F1-scores and the balanced precision and recall for both classes imply that the model is now perfectly calibrated and capable of dealing with the initially imbalanced dataset. The outcome highlights the necessity for utilizing the right resampling methods in dealing with class imbalance in classification problems.

Comparison of Classifiers

We employed two classifiers, Random Forest and Decision Tree, in the present study to compare and assess the performance of these models in determining the effectiveness of customer churn prediction. We tested each model under two conditions: using an imbalanced dataset and an imbalanced dataset that had been generated using the SMOTEENN sampling algorithm. Our results indicate that both classifiers demonstrated a notable boost in performance when trained with the balanced dataset, particularly in how well they were able to properly classify churners — our class of interest.

With the unbalanced dataset, overall accuracy of the Decision Tree classifier was 78% and F1-score of the churn class was 0.55, which shows poor identification of true churners. The Random Forest classifier was even slightly improved with an overall accuracy of 80% and an F1-score of the churn class being 0.57. Yet, the performance was still poor because the recall values for the churners were low it was difficult to handle the class imbalance.

Upon SMOTEENN application, a hybrid resampling strategy that integrates minority class oversampling and suspicious sample elimination, both classifiers had improved performance significantly. Decision Tree classifier achieved an excellent accuracy of 93%, precision of 91%, recall of 96%, and F1-score of 94% for the churn class. This reflects excellent balance in predicting real churners and avoiding false negatives.

Also, the Random Forest classifier with SMOTEENN dataset applied in training produced accuracy of 94%, precision of 93%, recall of 96%, and F1-score of 94% for churn class. These are indications that both classifiers, with balanced data, performed superbly in classifying churners.

Out of the two, Random Forest edges Decision Tree slightly in terms of both total accuracy and stability. However, the margin is thin, and both classifiers were equally good after application of SMOTEENN.

Performance of Churn prediction model

On our testing of Decision Tree and Random Forest classifiers on the imbalanced original telecom churn data, Decision Tree was medium overall accuracy but extremely low on the minority churn class. With no resampling, Decision Tree was 78% accurate (precision = 0.67, recall = 0.47, F1 = 0.55) whereas Random Forest was 80% accurate (precision = 0.74, recall = 0.47, F1 = 0.57). Especially, both classifiers performed very low recall (0.47) on the churn class – i.e., they identified fewer than half of true churners – even though they were fairly accurate. This is the characteristic behavior in imbalanced environments: the models will "be biased towards the majority class, having good accuracy but low recall for the minority class". That is, a classifier can classify nearly all customers as "non-churn" and remain seemingly correct, but catch all but a few churn cases. Due to the dominant size of the non-churn class, accuracy alone is in this case uninformative. Otherwise, we look at precision, recall, and F1 for a better picture:

- **Decision Tree (baseline):** Accuracy 78%, Precision 0.67, Recall 0.47, F1 0.55.
- **Random Forest (baseline):** Accuracy 80%, Precision 0.74, Recall 0.47, F1 0.57.

These figures indicate Random Forest to be slightly better than single tree overall on precision and accuracy, although both models' recalls were equally bad. In real life, this translates into over half of true churners being left behind by both models. This demonstrates the flaw of accuracy for imbalanced churn prediction: high accuracy can mask poor performance on minority churn class. Therefore, those measures like recall and the F1-score (which treats precision and recall as equal) provide a better assessment of churn detection.

Using SMOTE-ENN resampling also considerably enhanced performance for both models. Following class balancing using SMOTE-ENN (a combination of oversampling minority and removing noisy majority samples), the models became considerably more responsive to churn. The Decision Tree's Accuracy skyrocketed to 93% (Precision 0.91, Recall 0.96, F1 0.94), while the Random Forest achieved an accuracy of 93.8% (Precision 0.93, Recall 0.96, F1 0.94). Both models now have extremely high recall (0.96), meaning they mark 96% of all cases of churn correctly. The Random Forest has a slight edge over the Decision Tree in precision and accuracy but is otherwise identical once the data is balanced. Notably, F1-scores (0.94) are rising to almost perfect points, which means that the models are properly and balancedly identifying churners. The enhancements are owing to the effect of SMOTE-ENN: it "rebalances class distributions, enhancing the representativeness of minority classes and making the model more robust by reducing overfitting."

- **Decision Tree (with SMOTE-ENN):** Accuracy 93%, Precision 0.91, Recall 0.96, F1 0.94.
- **Random Forest (with SMOTE-ENN):** Accuracy 93.8%, Precision 0.93, Recall 0.96, F1 0.94.

Both classifiers are quite well-balanced in performance after resampling. Such high recall values are particularly useful: in churn modeling it is more essential to detect as many true churners as possible. In a B2B telecom environment, each lost customer account can be an excellent source of revenue loss, so missed detections (false negatives, undetected churners) are much more expensive than false alarms. As one applicable rule of thumb notes, "the RECALL number is more important than the Precision number for a churn algorithm". That is, it is more desirable to over-predict churn (and sacrifice precision) than to not predict future churners at all. Both models in these cases sacrifice some precision (~0.91–0.93) for practically flawless recall (0.96). The F1-score has the advantage of giving a single measure that balances between the two: in the context of imbalanced problems, the F1-score is a superior metric than accuracy. It only improves when the model correctly predicts more of the minority class instances, as our goal is to find all potential churners. From this comparison, some key observations are:

Accuracy is not sufficient in itself. A model can have high overall accuracy on imbalanced churn data by predicting the majority class predominantly. This is what happened to the baseline models. Accuracy and recall only became much better after resampling.

Recall is of major importance in churn prediction. Identification of true churners is of major importance in retention. We prefer recall over precision in this environment. False positives (non-churners predicted as churners) primarily result in additional marketing effort, while missed churners (unsuccessful churners) result in lost dollars and customers.

F1-score is regulated by precision and recall. In extremely imbalanced data, the F1-score is a great performance measure. Post-SMOTEENN F1 scores (≈0.94) show that the models are well choosing churners with little loss of accuracy.

Ensemble models perform reasonably. Random Forest, which is an ensemble of decision trees, will provide more robust, accurate results than one tree. Ensembles have been shown to "reduce bias and improve accuracy" in churn behaviors and to learn subtle data patterns better. This is borne out in our results: Random Forest slightly outperforms the individual tree, and they are both excellent after data balancing.

SMOTE-ENN works well. Combining oversampling and undersampling balances class imbalance by creating churn cases and removing noisy instances. SMOTE-ENN has been well-documented to improve minority-class prediction in churn and other usage. There it produced a drastic recall enhancement (from 0.47 to 0.96) for both classifiers, effectively eliminating the imbalance issue.

Briefly, performance test indicates that Decision Tree and Random Forest achieve good churn detection only after class imbalance has been tackled. Hybrid resampling method (SMOTE-ENN) and ensemble method (Random Forest) both perform extremely well: they convert median baseline models into high-sensitivity, accurate churn predictors. Most notably for telecom retention perhaps, we have maximized recall – detection of nearly all the churners – that aligns with business objectives of customer retention. The F1-scores also define the even-strength of the models. Overall, the comparison substantiates the weakness of accuracy in imbalanced datasets as well as the utility of precision-recall scores, particularly in churn situations where recall leads to the success of retention campaigns.

Implications

In the digitally and competitively expanding telecommunications environment of today, customer churn management is an important component of long-term business achievement. As part of a comprehensive customer relationship management (CRM) strategy, churn management consists of two fundamental activities: first, identifying which customers have the highest likelihood of churning, and second, enacting successful retention tactics—like loyalty programs or individualized interaction—to keep them from churning.

Retention programs are not just a reaction to losing customers but are value-added investment initiatives in order to capture the maximum lifetime value from customers. Effective customer retention can prevent having to constantly find new customers, and companies can concentrate on building relations with current users. Long-term loyal customers are less vulnerable to competition, yield stable revenue streams, and tend to become brand champions through word-of-mouth, all of which lead to increased profitability.

Lost customers impact the company in two folds: not only do they lower revenues, but the acquisition cost as well, which is always much greater than customer retention cost. Furthermore, the return on investment in retention initiatives is greater than on acquisition endeavors. Repeat purchases by current customers are also more probable, and more insensitive to price, and hence more worthwhile in the long run.

Thus, a precise churn prediction model is the key to ensuring that such efforts at retaining are not just cost-efficient and cost-saving. By precisely determining the customers who will churn, companies can address them with customized interventions rather than implementing blanket solutions for all the customers. Through this approached method, unnecessary resource spending on non-churn customers is avoided.

Within this project, our model of churn prediction has been able to reduce the occurrence of false positives and false negatives. This accuracy helps organizations make more informed decisions from the model's output, targeting real risks of churn with interventions and preventing wasting time and resources on customers who are not at risk.

Though the prediction model itself is just part of the solution, its real worth is when coupled with clever retention tactics attuned to the at-risk clients uncovered. This is particularly critical in repeat revenue subscription-based telecom services B2C and B2B, where loyalty of customers and wellness of a stable customer base for repeat business hinge so much on repeated revenues. Lower churn rates are directly tied to improved bottom lines and increased customer lifetime value.

With that being the case, despite the best predictive models, churn cannot be done away with completely. Customers will still churn out as a result of any of countless uncontrollable variables. Churn is either voluntary—where a customer willingly decides to churn—or involuntary, e.g., for reasons of non-payment or system issues. Our model currently does not differentiate between these but even the identification of would-be churners improves the capacity of a firm to actively re-engage and retain at least some of them.

Each retained customer makes a contribution to the bottom line. Incremental retention gains can yield massive profitability gains. That makes it an essential ability for telecom operators to combine data science-based churn prediction and successful customer retention tactics in order to continue to grow and retain satisfied customers in an extremely competitive market.

Conclusion

Research Question: We have demonstrated in this study that supervised machine learning can effectively predict telecom customer churn. By reformulating churn prediction as a classification task, we demonstrated that classifiers learned from past customer data can determine high-risk accounts to churn. Consistent with prior work, our Decision Tree and Random Forest models were correct in predicting churners, thereby establishing the applicability of ML-based churn prediction.

Classifier Performance: Decision Tree and Random Forest both showed very strong predictive performance on the churn data. Interestingly, Random Forest scored slightly better than the individual decision tree on all the key metrics (accuracy, precision, recall, F1-score), as would be the case with ensemble techniques. This finding is consistent with previous findings that ensemble learners such as Random Forest tend to be more accurate than individual trees. The marginal advantage of Random Forest over precision, recall, and F1-score explains its stability for churn classification under this condition.

Handling Data Unbalance: The initial churn data were grossly unbalanced (many fewer churners). We needed to balance classes with the help of SMOTEENN (a combination oversampling-cleaning approach). Not surprisingly, SMOTEENN greatly enhanced the recall (sensitivity) of the churn prediction (culling more actual churners) at a minimal sacrifice in precision. That is, the model improved in culling customers at risk with an intermediate increase in false alarms. This is consistent with conventional practice: oversampling the minority class tends to improve recall but reduce precision. Overall, the SMOTEENN step improved the overall F1 score and provided model predictions with improved class balance, validating the effectiveness of sophisticated resampling in churn tasks

REFERNCES

1. Ahmad, A., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data, 6*(28), 1–24. https://doi.org/10.1186/s40537-019-0191-6
2. Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3), 4626–4636. https://doi.org/10.1016/j.eswa.2008.05.027
3. IBM. (2020). *Predicting customer churn: A case study using IBM Watson Studio*. IBM Developer. https://developer.ibm.com/articles/predicting-customer-churn/
4. Towards Data Science. (2019, June 15). *Customer churn prediction in Python*. https://towardsdatascience.com/customer-churn-prediction-in-python-d2d0543eb3b0
5. Analytics Vidhya. (2022, February 7). *A complete guide to churn prediction with machine learning*. https://www.analyticsvidhya.com/blog/2022/02/a-complete-guide-to-churn-prediction-with-machine-learning/
6. Turing. (2023). *How machine learning can reduce customer churn for telecom companies*. https://www.turing.com/blog/how-machine-learning-can-reduce-customer-churn-in-telecom/
7. DataCamp. (2021). *Telecom churn prediction project*. https://www.datacamp.com/projects/557
8. KDnuggets. (2020). *10 best practices for customer churn prediction*. https://www.kdnuggets.com/2020/08/10-best-practices-customer-churn-prediction.html
9. Oracle. (2021). Strategies for reducing telecom churn. https://www.oracle.com/industries/communications/solutions/customer-churn-reduction/

11% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 10 words)

Match Groups

69 Not Cited or Quoted 11%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

7%	Internet sources
4%	Publications
8%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **69** Not Cited or Quoted 11%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

7%	Internet sources
4%	Publications
8%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	dspace.dtu.ac.in:8080	2%
2	Internet	www.duo.uio.no	<1%
3	Internet	www.mdpi.com	<1%
4	Publication	de Sousa, Lúcia Maria Bessa. "Detecting a Poker Face", Universidade de Aveiro (P...	<1%
5	Submitted works	Gisma University of Applied Sciences GmbH on 2025-03-27	<1%
6	Submitted works	University of North Texas on 2024-10-05	<1%
7	Submitted works	University of Salford on 2022-05-16	<1%
8	Internet	ijisrt.com	<1%
9	Submitted works	University of Hertfordshire on 2024-12-12	<1%
10	Internet	www.univio.com	<1%