ANALYSING TELECOM CUSTOMER CHURN

PREDICTIVE MODELLING AND RETENTION STRATEGIES

Laxmi Shashank Poralla
Data Analytics and Engineering
George Mason University
Fairfax, USA
lporalla@gmu.edu

Sesha Sai Dheeraj Mantripragada
Information Systems
George Mason University
Fairfax, USA
smantri4@gmu.edu

Lalitha Dandibhotla
Data Analytics and Engineering
George Mason University
Fairfax, USA
ldandih@gmu.edu

***Abstract*—The telecommunications industry faces significant challenges with customer churn, which substantially impacts profitability due to high customer acquisition costs and lost revenue from defections. This paper presents a comprehensive study on predictive modeling and retention strategies to mitigate churn in the telecom sector. Through detailed analysis leveraging advanced data analytics and machine learning techniques, we identify critical predictors of churn and examine customer behavior patterns. The study integrates various predictive models, including logistic regression, random forests, and gradient-boosted trees, to enhance the accuracy of churn prediction. By profiling high-risk subscribers and implementing targeted retention programs, telecom providers can potentially extend customer tenure and improve overall business sustainability. The outcomes of this research offer actionable insights into designing effective churn prevention initiatives, thereby contributing to both academic discourse and practical applications within the industry.**

Keywords—Telecommunications, Customer Churn, Predictive Modeling, Retention Strategies, Data Analytics, Machine Learning

I. INTRODUCTION

Customer churn, or subscribers quitting services, is a major worry for providers in the fiercely competitive telecommunications industry. The telecom industry has extremely narrow profit margins and depends heavily on long-term subscriber retention to provide value for investors. As a result, telecoms immediately lose money equal to the monthly rates paid by those particular clients when they disconnect. Furthermore, the industry incurs substantial acquisition expenses, such as advertising promotions, device deals, and retail marketing, to replace the departed subscribers. Studies have revealed that acquiring a new customer can cost five to ten times more than simply retaining an existing one. Additionally, long-tenured customers provide extended value through recommendations, family plan sign ups, and the adoption of new offerings, resulting in a total lifetime value assessed at two to five times higher than the average revenue per user (ARPU).

The churn rates across telecom carriers typically range from 1% to 3% each month, translating to an annual customer turnover of 12% to 36%. With over a million customers, this level of churn can result in the loss of up to 360,000 subscribers annually. Considering an average ARPU of $50 to $80 per month and a monthly churn rate of 2%, telecoms forfeit approximately $1 million in revenue each month due to defecting customers. Consequently, successfully reducing the churn rate can enhance profitability by 25% to 125%, translating to billions in savings if churn is mitigated effectively.

Furthermore, operations related to customer assistance and account closure are associated with discontinuing customers. Hence, being able to determine subscribers prone to quit in advance allows implementing customized loyalty incentives or addressing their issues proactively. While adding new consumers spurs growth, curbing churn establishes a foundation for organic expansion through cross-sales, renewed contracts, and positive word-of-mouth. Studies have revealed that retaining customers has over 600% higher returns compared to continually expanding resources to acquire replacements.

By constructing a robust predictive model and profiling the characteristics of high-risk subscribers, telecom providers can shape targeted retention programs, personalized marketing offers, and tailored customer communications to effectively reduce subscriber loss. The end goal is to establish durable relationships with customers, minimize sensitivity to temporary service disruptions or pricing changes, and foster long-term loyalty through data-driven churn prevention initiatives.

II. MOTIVATION

Many strong elements present in the telecommunications sector as well as modern business needs are the motivations. First off, loss of customers is a major issue for the telecommunications industry that has a significant impact on the viability and profitability of businesses. Because of the industry's natural competitiveness, as well as the increasing pace of technical development and changing consumer tastes, it is imperative that churn management be given strategic emphasis. The project aims to tackle a basic issue that has a direct influence on telecom firms' profits by exploring this field.

Additionally, the project's objectives are in line with more general developments in predictive analytics and data-driven decision-making across industries. Data has become a crucial component of well-informed decision-making in the current digital era, providing organizations with previous unheard-of chances to extract useful insights from big datasets. Through the application of predictive modelling approaches, the project aims to analyze telecom customer churn and

leverage data analytics to inform strategic actions that improve customer retention and promote long-term commercial success.

Additionally, the initiative is highly relevant from an academic standpoint, providing a chance to add to the corpus of information already available on customer churn prediction and retention tactics in the telecom industry. The initiative intends to produce insights that benefit individual telecom companies as well as educate industry-wide best practices and academic discourse through rigorous empirical research and analysis.

In conclusion, the motivation is supported by the research's scholarly value, possible societal impact, alignment with current data analytics developments, and strategic importance within the telecoms sector. The project aims to make a significant contribution to academic research and real-world commercial solutions in the ever-changing telecoms industry by tackling this important topic.

III. Literature Review

Significant study has been conducted in the area of customer churn prediction as a result of the telecom industry's changing landscape. This is an important topic for customer retention and long-term business growth. In order to manage customer churn with advanced data analytics and machine learning approaches, this literature review looks at key studies in the field, each of which offers distinctive insights and methodologies.

A. Customer churn prediction in telecommunications

Huang, Kechadi, and Buckley (2012) present a thorough collection of features in their first study that can be used to forecast customer turnover in landline telecommunication services. Using seven predictive modelling methodologies, their innovative approach shows how incorporating a wide range of data points—such as billing, call details, and customer complaints—can effectively improve the accuracy of churn prediction. This study shows an improvement in predictive skills over existing models and establishes the foundation for using extensive customer interaction and behaviour data to anticipate turnover.[1]

B. A Customer Churn Prediction Model in Telecom Industry Using Boosting

A later study investigated the use of boosting algorithms to improve churn prediction models. By grouping customers into clusters according to churn risk, boosting was used in this study to improve model performance, which focused on customer turnover in the telecom sector. A sophisticated method of detecting high-risk clients was made possible by the creative clustering based on boosting algorithm weights and the use of logistic regression as the base learner. This methodological breakthrough provided strategic insights into customer segmentation for focused retention campaigns in addition to increasing prediction accuracy. [2]

C. Customer Churn Prediction in Telecommunication Industry Using Deep Learning

In the third important study, which focused on deep learning, customer turnover was predicted with remarkable accuracy using a Deep Backpropagation Artificial Neural Network (Deep-BP-ANN). Through the use of two feature selection techniques, Variance Thresholding and Lasso Regression, in conjunction with overfitting prevention methods like dropout and activity regularization, this study showed how much better deep learning models are able to handle large datasets and intricate feature relationships that are typical in the telecom industry. The ability of deep learning to uncover meaningful patterns from large amounts of customer data was demonstrated by the Deep-BP-ANN model's performance in beating traditional machine learning techniques. This opens up a promising route for future churn prediction attempts.[3]

D. Customer Churn Prediction for Telecom Services

The study by Yabas et al., delves into the realm of churn prediction by leveraging machine learning techniques to analyze customer behavior and service usage patterns. Their study employs various algorithms to forecast the likelihood of customers discontinuing their services. The authors emphasize the importance of understanding the underlying factors contributing to churn, such as service dissatisfaction or competitive offerings, which can significantly aid in developing tailored customer retention strategies.[4]

E. Telecom Customer Churn Prediction: A Survey

The fifth study provides a comprehensive review of the churn prediction landscape in telecommunications. They dissect the evolution of churn prediction methods, from traditional statistical models to advanced machine learning techniques, highlighting the transition towards data-driven decision-making in the telecom industry. The paper serves as a vital resource for researchers and practitioners, offering a detailed analysis of the strengths and limitations of various churn prediction models and their applicability in real-world scenarios.[5]

F. Machine Learning Based Telecom-Customer Churn Prediction

The next study takes a deep dive into the application of machine learning for churn prediction. By comparing the effectiveness of different machine learning models, the study sheds light on how these technologies can be optimized to predict customer churn accurately. The authors' analysis reveals critical insights into the behavior patterns of customers who are likely to churn, providing valuable information for telecom companies to refine their customer retention strategies.[6]

G. Customer Churn Prediction in Telecommunication: An Analysis on Issues, Techniques and Future Trends

The final study addresses the broader issues and trends in churn prediction research. The paper critiques the current state of churn prediction methods and discusses the emerging trends that are shaping the future of churn analytics in the telecom sector. Their findings underscore the need for ongoing research and development in this area to enhance the accuracy and reliability of churn prediction models.[7]

H. *Customer Churn Prediction Using Machine Learning and Information Gain Filter Feature Selection*

Saheed and Hambali propose a churn prediction model for telecom using SVM, MLP, RF, and NB, with Information Gain and Ranker feature selection. With 10-fold cross-validation, the model achieved 95.02% accuracy with feature selection, surpassing 92.92% without. This method enhances predictive accuracy, aiding targeted retention strategies in high-churn industries.[8]

I. *Churn Prediction in Telecommunication Industry Using Machine Learning Techniques*

This study by Yashraj Bharambe, Pranav Deshmukh, Pranav Karanjawane, Diptesh Chaudhari, and Dr. Nihar M. Ranjan presents a comprehensive model for predicting customer churn in the telecommunications industry. The research employs various machine learning techniques, including Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost, to develop a predictive model capable of identifying high-risk churn customers. By leveraging historical data, the model accurately assesses the likelihood of customer defection, thereby enabling telecom companies to implement targeted retention strategies more effectively. The study highlights the critical role of machine learning in enhancing the accuracy of churn prediction models and suggests that these advanced analytical techniques can significantly reduce customer attrition, ultimately benefiting the company's revenue and customer service standards. [9]

J. *Advanced Churn Prediction Models in Telecommunications Using Diverse Machine Learning Techniques*

Abdul Razak and Wahid examine telecom churn prediction using various ML models, emphasizing usage patterns. Their study compares Linear Regression, Random Forest, SVM, KNN, and Decision Tree, with Random Forest showing the highest accuracy at 95.5%. It highlights the efficacy of advanced ML techniques in churn prediction, providing insights into model performance across multiple metrics. While no single model dominates universally, Random Forest emerges as a robust choice for enhancing customer retention strategies.[10]

K. *Predicting customer churn prediction in telecom sector using various machine learning techniques*

Gaur and Dubey explore ML techniques for telecom churn prediction, crucial for loyalty and profitability. They employ Logistic Regression, SVM, Random Forest, and Gradient Boosting Trees, with Gradient Boosting Trees leading with an AUC of 84.57%. The study underscores the significance of accurate churn prediction for targeted retention strategies in telecom. It showcases the potential of advanced analytics in strategic business applications, aiding in customer retention and service enhancement.[11]

L. *Predicting customer churn prediction in telecom sector using various machine learning techniques*

This paper by Ashish Sharma and his team addresses the challenge of customer retention in the rapidly evolving telecom sector within smart cities. Utilizing advanced machine learning techniques, specifically Decision Trees and Logistic Regression, the study analyzes a dataset of 3,334 instances to predict customer churn. The models developed aim to discern patterns that indicate likely customer departure, which is crucial for maintaining a robust customer base in competitive markets. The Decision Tree model demonstrated a remarkable accuracy of 97%, while the Logistic Regression model achieved an accuracy of 80%. These findings underscore the potential of machine learning to enhance predictive accuracy in churn management, thereby aiding telecom companies in devising more effective customer retention strategies and reducing the high costs associated with acquiring new customers.

IV. PROPOSED APPROACH

A. *Data Processing*

The raw customer data will go through a thorough cleaning and preprocessing phase before analysis and modeling can begin. Any missing data points will be handled using techniques like mean/median imputation for numerical variables and mode imputation for categorical ones. Rows with an excessively high percentage of missing values may need to be removed entirely. Categorical features like contract type, payment method, etc. will be encoded into numerical representations that the models can understand. Highly skewed numerical variables like monthly charges or data usage will be normalized using methods like log transformation or box-cox to ensure the distributions are closer to normal. The "Tenure" variable measuring a customer's time with the company will be binned into tenure cohorts like 0-6 months, 6-12 months, etc. to better study churn patterns over time. Finally, the data will be split into training and test partitions using stratified random sampling. This ensures the distributions of churners vs non-churners are preserved proportionally in both sets, preventing any leakage.

B. *Exploratory Analysis*

Extensive univariate and bivariate analysis techniques will be leveraged to uncover patterns, trends, and correlations between the input variables and the target churn variable. This involves visualizations like histograms, density plots, box plots for numerical variables and bar charts for categorical ones across churners and non-churners. Correlation matrices and scatter plots will identify the strength and nature (linear/non-linear) of relationships between different variables and churn propensity. Variables with very high correlation may need to be removed or combined to reduce multicollinearity issues. The analysis may uncover particularly high-risk customer segments that are more susceptible to churn based on combinations of demographics, service subscriptions, usage levels etc. These segments would need to be thoroughly profiled to extract insights around what factors are driving their elevated defection rates. If needed, new features may be engineered specifically capturing such high-risk groups. Interactive graphs will also be built to visualize geographic differences in churn patterns across states/cities. These visualizations can reveal region-specific challenges, identify critical periods of high churn risk

over the customer tenure lifecycle, and quantify how additional adoption affects churn rates.

C. Model Building

Based on the exploratory analysis, the most predictive features related to capturing engagement levels, lifetime value potential, and general demographics/service usage will be consolidated into composite scores or factors. Dimensionality reduction techniques like PCA may be used to extract these high signal components. These key derived indicators will then be used to model churn status (stayed vs churned) through multiple machine learning classification algorithms. Benchmark linear models like logistic regression will be trained. Tree-based ensemble methods like random forests and gradient boosted trees can capture more complex non-linear relationships. K-fold stratified cross-validation will be used, retraining the algorithms k times across different continuous partitions of the training set.

D. Model Evaluation

Standard classification metrics like overall accuracy, precision, recall, F1 scores calculated against the held-out test set will be used to evaluate the predictive performance of the models. Given the likely class imbalance with many more non-churners, the focus will remain on the precision and recall values specifically for the "churned" class. Techniques like ROC curves and area under the curve (AUC) metrics will assess the models' ability to discriminate between churners and non-churners at various probability thresholds. Precision-recall curves will also identify the tradeoffs between the two metrics at different decision boundaries. The most accurate, robust model will be selected to operationalize churn risk scoring on new customers. It can stratify subscribers into deciles based on their predicted churn probabilities. Detailed profiles of the highest risk decile can uncover vulnerabilities to strategically design and prioritize retention campaigns and incentives. Model monitoring processes will be established to track the champion model's performance over time on new data. If degradation is observed, retraining will be needed to maintain effectiveness.

V. Data Analysis and Preprocessing

To commence the analysis, the raw dataset underwent manual inspection using spreadsheet software to review the data descriptions and become familiar with the variables. Following this the data underwent a thorough cleaning process to address missing values, outliers, and inconsistencies. This involved techniques such as imputation for missing data, removal of duplicate entries, and dimensionality reduction where necessary. Subsequently, exploratory visualizations and statistical techniques were employed to uncover preliminary insights regarding the factors influencing customer churn behavior.

A. Data Cleaning

In the data cleaning phase, our primary focus was on rectifying missing values in the dataset. Employing a variety of statistical methods and visualization aids, we meticulously identified areas with missing data. Figure 1, a bar chart, vividly showcases the distribution of these missing values across different variables. This visualization was pivotal in prioritizing our attention towards the most affected variables.

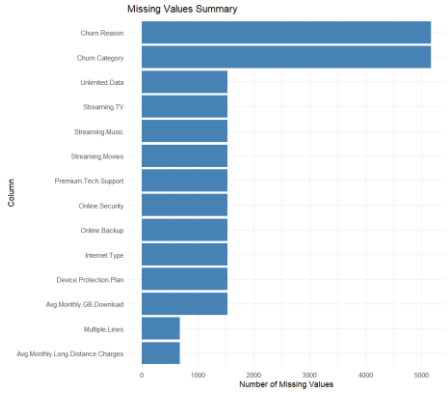

Figure 1: Missing Values Summary

Upon closer inspection, it became evident that all missing values were denoted as strings. To streamline the treatment of such instances and to ease subsequent analysis, a unanimous decision was made to substitute these missing values with a universal placeholder, namely 'N/A'. This approach ensures consistency in handling missing data while retaining incomplete records for analysis purposes. Moreover, opting for imputation over row deletion mitigates the risk of data loss and skewed outcomes. With missing values addressed, our next task was to scrutinize the dataset for duplicate entries. Fortunately, after a thorough examination, it was confirmed that the dataset harboured no duplicate records, alleviating the need for further action in this domain.

Having tackled missing values and duplicates, we proceeded to transform categorical columns into numerical representations utilizing one-hot encoding techniques. This transformation not only facilitates the inclusion of categorical variables in subsequent analyses but also aids in uncovering underlying patterns and relationships within the data. By converting categorical variables into numerical equivalents, we ensure compatibility with analytical tools and techniques employed in subsequent phases.

Through our concerted efforts in addressing missing values, removing duplicates, and converting categorical variables, the dataset underwent crucial preparatory steps for deeper analysis and interpretation. This methodical approach to data cleaning serves to mitigate potential biases and errors, thus bolstering the validity and robustness of ensuing analyses.

B. Statistical Analysis

1. Correlation Analysis

A correlation heatmap was constructed to visualize the correlation coefficients between the various customer-related metrics. Figure 2 depicts the relationships discussed. Age positively correlates with dependents, suggesting life-stage impacts on service needs.

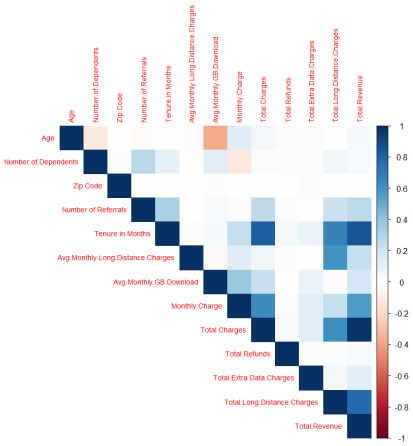

Figure 2: Correlation Analysis

Total charges strongly negatively corelates with refunds, implying higher charges lead to fewer refunds. Monthly charges positively correlate with usage factors, while tenure highly positively correlates with total charges, indicating longer-term customers generate more revenue. Surprisingly, zip codes show little correlation, suggesting location minimally impacts metrics considered. Referrals slightly positively correlate with tenure hinting at loyal customers engaging referral programs. Average monthly data usage weakly positively correlates with monthly charges, indicating higher usage marginally increases charges.

In conclusion, it's crucial to bear in mind that correlation does not imply causation, and there may be other underlying factors influencing the relationships observed in the data. Moreover, correlation coefficients solely assess linear relationships, potentially overlooking more intricate associations between variables. Therefore, it is imperative to conduct further analysis and leverage domain knowledge to accurately interpret the results and draw meaningful conclusions.

2. Descriptive Statistics

To investigate the impact of contractual obligations on customer churn, churn rates were stratified and visualized across different contract types: month-to-month, one-year, and two-year contracts. Figure 3 depicts the relationships discussed. The stacked bar chart clearly illustrated that customers with month-to-month contracts exhibited substantially higher churn rates (37%) compared to those with longer-term annual (18%) or bi-annual (12%) commitments. This finding suggests that customers with shorter contractual periods are more susceptible to discontinuing services.

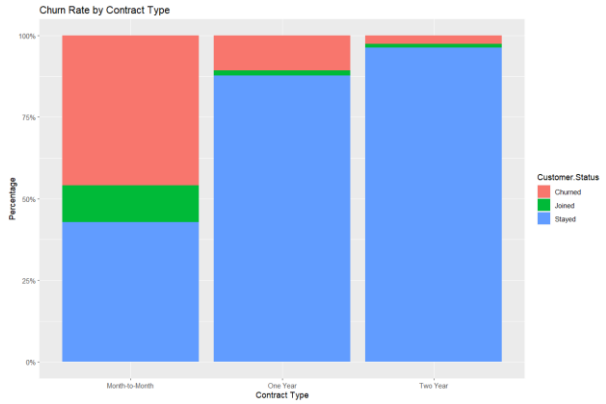

Figure 3 : Churn Rate by Contract Type

A histogram was generated to examine the distribution of customer tenure, differentiated by their current status (churned, joined, or stayed). Figure 4 depicts the relationships discussed. The visualization revealed a notable trend wherein the majority of churn events occurred among customers with relatively shorter tenures with the company. This observation indicates that newer subscribers may be at an elevated risk of churning compared to their more tenured counterparts.

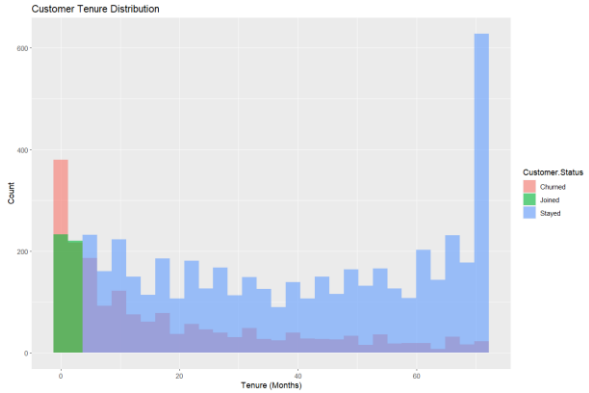

Figure 4 : Customer Tenure Distribution

To gain insights into the underlying factors driving customer churn, a tree map and pie chart were constructed to analyze the stated reasons for churn, categorized into broader groups. Figure 5 and Figure 6 depicts the chart and treemap respectively. The tree map highlighted "Competitor" as the dominant category, occupying the largest area, while the pie chart quantified this category as accounting for 45% of all churn reasons. This finding suggests that a substantial proportion of customers are leaving for better offers, services, or perceived value from rival providers. Other prominent categories included issues related to attitude/service quality, pricing complaints, and lack of desired features or product capabilities.

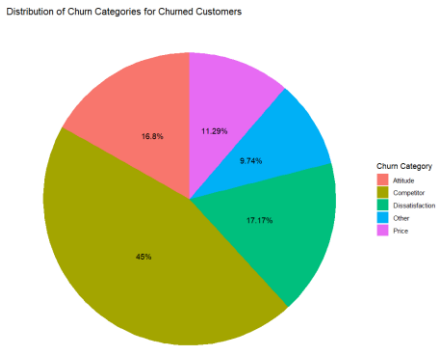

Figure 5 : Distribution of Churn Categories for Churned Customers

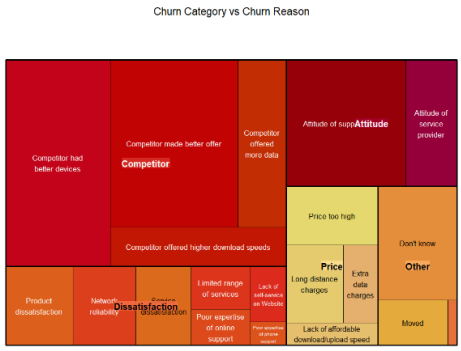

Figure 6 : Churn Category vs Churn Reason

To assess the potential impact of marketing offers on customer retention, a bar chart was created to visualize customer status (churned, joined, or stayed) across different offer types. Figure 7 depicts the relationships discussed. The chart revealed that a significant number of customers who ultimately churned did not receive any offers. This observation implies that providing targeted retention offers could potentially aid in mitigating customer attrition rates.

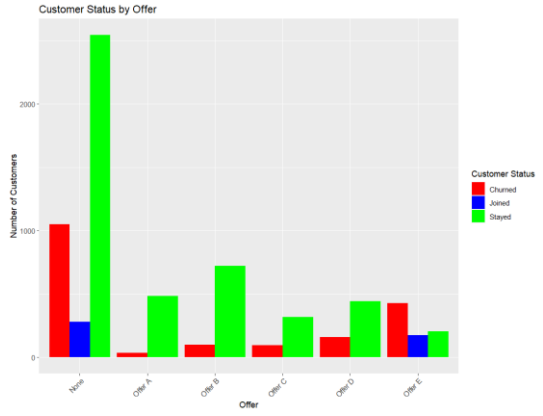

Figure 7 : Customer Status by Offer

A geomap was generated to examine the spatial distribution of churned customers across different geographic regions. Figure 8 depicts the relationships discussed. The visualization uncovered specific areas with higher concentrations of red dots, indicating locations that experienced elevated levels of customer churn compared to others. However, the results were not as effective as anticipated and did not provide significant insight into the underlying factors contributing to churn.

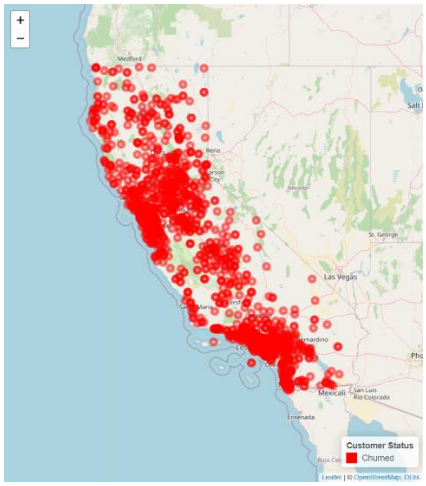

Figure 8 : Geomap of Customers Churned

These initial discoveries provided the basis for more thorough data preprocessing, feature engineering, and the creation of reliable prediction models that correctly predicted clients who were at a high risk of leaving. Furthermore, comprehending the main factors that were discovered during this initial analysis stage enabled the development and execution of targeted interventions and client retention plans that effectively reduce attrition.

3. Encoding

In the described methodology, a systematic approach to preprocessing a dataset for analytical readiness, specifically focusing on the conversion of categorical variables to a numerical format, is delineated. This conversion is crucial for deploying statistical models and machine learning algorithms that necessitate numerical input for optimal performance. The initial step involves the identification of categorical variables within a dataset, which are typically characterized by non-numeric, discrete values representing various categories or classes associated with the data attributes. Common examples of such attributes include demographic information like gender or service options such as types of internet subscriptions.

Following the identification process, each unique category within these variables is assigned a distinct numeric code. This assignment is pivotal as it transforms qualitative data into a quantitatively interpretable format, thereby facilitating the application of numerical operations and algorithms. The mapping from categorical labels to numeric codes is meticulously preserved, typically in a structured format, to ensure transparency and reversibility of the encoding process. This mapping is essential not only for the interpretability of the data by analysts and stakeholders but also for maintaining the integrity of data through subsequent processing stages.

The encoded data, now transformed, is integrated back into the original dataset, replacing the categorical labels with their respective numeric codes. This encoded dataset represents a transformed version of the original data, formatted to support various analytical procedures that require numeric inputs. To support reproducibility and further analysis, the mappings of labels to numeric codes are saved separately, often in a dedicated file. This practice ensures that any future interpretations or analyses can reference the original categorical contexts accurately.

In summary, this methodology underpins the preparatory phase of data analysis, where data in its raw categorical form is systematically transformed into a structured numeric format. This transformation is critical for leveraging advanced statistical techniques and machine learning models, which require numeric input to function effectively. The process not only enhances the usability of the dataset but also contributes to the robustness and validity of the analytical outcomes derived from it.

VI. Model Building

A. Data Preprocessing

Prior to commencing model training and evaluation, comprehensive data preprocessing procedures were undertaken to ensure the quality and relevance of input features. A pivotal aspect of this preprocessing involved meticulous selection of input variables for inclusion in the modeling phase. Given the telecom dataset's nature, initially comprising 7043 rows and 32 columns, it was imperative to discern and eliminate columns that might not significantly contribute to the predictive task at hand.

Consequently, columns such as 'Customer.ID', 'City', 'Latitude', 'Longitude', and 'Churn.Reason' were identified and removed. These columns were deemed either irrelevant or redundant for predicting customer churn or customer status. The elimination of these columns not only reduced dimensionality but also streamlined the modeling process by concentrating on the most informative features.

B. Model Selection

In our study, we undertook an extensive examination of various machine learning methodologies aimed at forecasting customer behavior and churn within the telecom sector. The array of models scrutinized encompassed Logistic Regression, Random Forest, Gradient Boosting, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Naive Bayes, and Artificial Neural Networks (ANN). Each of these models presents distinct advantages and applicability to different data types and prediction objectives. Logistic Regression, for instance, represents a straightforward yet potent linear model frequently employed in binary classification tasks. Meanwhile, Random Forest and Gradient Boosting stand as ensemble techniques amalgamating multiple decision trees to enhance predictive accuracy. SVM proves effective for both linear and non-linear classification tasks, whereas KNN relies on data point similarity for predictions. Naive Bayes operates as a probabilistic classifier grounded in Bayes' theorem, and ANN emerges as a versatile model inspired by human brain functionality. By embracing a diverse array of models, our aim was to pinpoint the most suitable approach for forecasting customer churn within the telecom dataset.

C. Feature Engineering

Feature engineering constitutes a pivotal aspect in constructing precise and resilient predictive models. Throughout our analysis, we employed various feature engineering methodologies to preprocess the dataset and amplify the predictive efficacy of the models. This included one-hot encoding for categorical variables, transforming categorical attributes into binary vectors to ensure compatibility with machine learning algorithms. Furthermore, we applied scaling or normalization to numerical attributes to standardize their range and forestall biases during model training. Additionally, we delved into the creation of novel features from existing ones through techniques like feature extraction and transformation. By meticulously engineering the features, our objective was to encapsulate meaningful patterns and correlations within the data, thereby facilitating more effective churn prediction.

D. Train-Test Split

The division of the dataset into training and test sets via a stratified random sampling technique served as a crucial step in assessing model performance. This stratification guaranteed the preservation of target class distribution (e.g., churned vs. joined vs. stayed) in both the training and test sets. With an initial dataset size of 7043 rows, 70% of the data was allocated for training purposes, while 30% was reserved for testing. By stratifying the data based on the target variable, we mitigated the risk of biased model evaluation and obtained more dependable estimates of predictive accuracy.

E. Model Training

Upon the partitioning of the dataset into training and test sets, we proceeded to train each selected model utilizing the scaled training data. Leveraging the scikit-learn library in Python, which offers an extensive suite of machine learning algorithms and tools for model training and evaluation, we followed a uniform workflow for each model. This involved fitting the model to the training data and fine-tuning its hyperparameters to optimize performance. Through iterative adjustments of the model parameters and assessment of performance on the validation set, we endeavored to identify the optimal configuration for each model. This systematic approach enabled us to construct robust predictive models proficient in accurately identifying potential churners within the telecom dataset.

VII. Model Evaluation

A. Evaluation metrics

In our model evaluation process, we employed a variety of evaluation metrics to assess the performance of the machine learning models trained on the telecom dataset. These metrics provide insights into different aspects of model performance,

allowing us to gauge their effectiveness in predicting customer behavior and churn. Key metrics utilized include accuracy, precision, recall, and F1-score, each offering unique perspectives on the model's predictive capabilities. Accuracy measures the overall correctness of predictions, while precision quantifies the proportion of true positive predictions among all positive predictions. Recall, also known as sensitivity, evaluates the ability of the model to correctly identify positive instances out of all actual positives. F1-score, which combines precision and recall into a single metric, provides a balanced assessment of the model's performance, particularly useful in scenarios with imbalanced class distributions.

The Table 1 and Figure 9 encapsulates the performance metrics of various machine learning models employed for telecom churn prediction, indicating their accuracy, precision, recall, and F1-score. Each model underwent rigorous evaluation, with logistic regression yielding an accuracy of 81.83% and demonstrating commendable precision, recall, and F1-score at 82.13%, 81.83%, and 81.96%, respectively. Random forest exhibited improved accuracy at 83.63% while maintaining competitive precision, recall, and F1-score metrics at 82.90%, 83.63%, and 83.11%, respectively. Gradient boosting surpassed previous models with an accuracy of 85.05% and displayed notable precision, recall, and F1-score at 84.47%, 85.05%, and 84.63%, respectively. SVM achieved an accuracy of 80.27% alongside precision, recall, and F1-score measures at 79.62%, 80.27%, and 79.65%, respectively. However, k-nearest neighbors and naive Bayes models showcased comparatively lower accuracies at 74.21% and 71.13%, with modest precision, recall, and F1-scores. Finally, the artificial neural network model attained an accuracy of 79.98% and exhibited consistent performance across precision, recall, and F1-score metrics at 79.92%, 79.98%, and 79.92%, respectively. This comprehensive evaluation provides valuable insights into the predictive capabilities of each model, facilitating informed decision-making in telecom churn prediction endeavors.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	81.83%	82.13%	81.83%	81.96%
Random Forest	83.63%	82.90%	83.63%	83.11%
Gradient Boosting	85.05%	84.47%	85.05%	84.63%
SVM	80.27%	79.62%	80.27%	79.65%
k-Nearest Neighbors	74.21%	74.37%	74.21%	74.08%
Naive Bayes	71.13%	77.12%	71.13%	72.81%
Artificial Neural Network	79.98%	79.92%	79.98%	79.92%

Table 1: Performance Overview for 'Customer.Status'

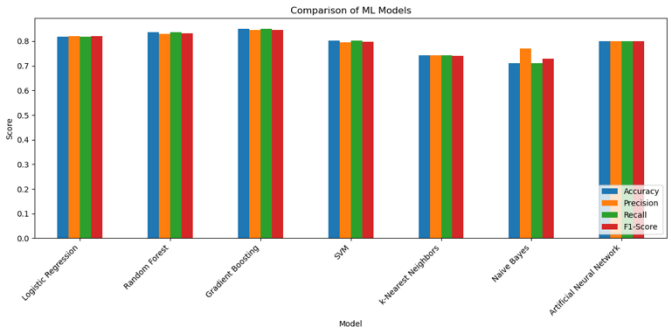

Figure 9 : Comparative Analysis of ML models for 'Customer.Status'

B. Best Model Selection

In selecting the best model from the candidate pool, we employed a rigorous evaluation process based on the custom score, which considers accuracy, precision, recall, and F1-score simultaneously. By averaging these metrics, we obtained a comprehensive measure of overall model performance, allowing for fair comparison and selection of the most suitable model for our predictive tasks. The chosen model demonstrated superior performance across multiple evaluation criteria, indicating robust predictive capabilities and suitability for real-world applications. This selection criterion ensured that the chosen model met the desired performance standards and could effectively address the predictive challenges posed by customer churn in the telecom industry.

Among the array of models evaluated, the Gradient Boosting model emerged triumphant as the optimal choice for our predictive tasks, boasting an impressive performance across various metrics. With an accuracy of 85.05%, precision of 84.47%, recall of 85.05%, and F1-Score of 84.63%, it exhibited robust predictive capabilities, particularly crucial in addressing the challenges posed by customer churn within the telecom industry. This superiority was further underscored by its confusion matrix. Notably, the model correctly identified 365 instances of churned customers, 98 instances of joined customers, and 1334 instances of customers who stayed. Such precision in prediction, validated through rigorous evaluation, signifies the model's efficacy in delineating and mitigating churn risks, thus rendering it a formidable asset in real-world telecom applications.

C. Confusion Matrix

To gain deeper insights into the performance of the best-performing model, we visualized the predictions using confusion matrices. The matrix provides a detailed breakdown of correct and incorrect predictions for each class, enabling us to assess the model's strengths and weaknesses across different categories. The confusion matrix represents the classification results for the three classes: "Churned", "Joined", and "Stayed". Each cell in the confusion matrix represents the count of true positive, false positive, true negative, and false negative predictions for a particular class. By plotting the confusion matrices for the best-performing model, we were able to identify patterns of misclassification and assess the model's ability to accurately predict specific outcomes. This visual

representation served as a valuable tool for understanding the model's predictive behavior and informing potential improvements.

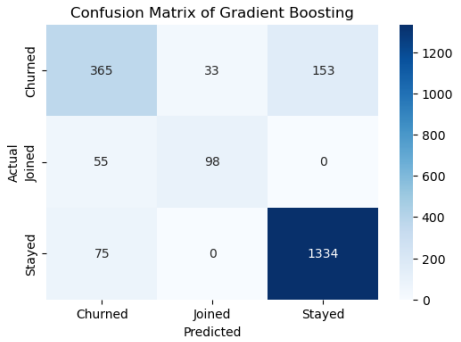

Figure 10: Confusion matrix of Gradient Boosting

The confusion matrix in Figure 12 depicts the performance for the best model i.e., gradient boosting model applied to a telecom churn dataset. The matrix reveals that the model correctly predicted 365 instances of customers who churned, while misclassifying 33 churned customers as joined and 153 as stayed. Additionally, it correctly identified 98 customers who joined without erroneously classifying any as churned or stayed. However, it failed to predict any customers who actually stayed in the network, misclassifying 75 as churned. Remarkably, it accurately predicted a vast majority of stayed customers, totaling 1334 instances. This comprehensive evaluation underscores the model's effectiveness in identifying churned and joined customers but highlights a slight limitation in accurately classifying stayed customers, suggesting potential avenues for model refinement.

D. ROC Curve

In addition to confusion matrices, we utilized Receiver Operating Characteristic (ROC) curves to evaluate the discrimination ability of the best-performing model. ROC curves plot the true positive rate (sensitivity) against the false positive rate (1-specificity) across different threshold values, providing a comprehensive visualization of the model's performance across various classification thresholds. We calculated the Area Under the Curve (AUC), which quantifies the overall performance of the model in distinguishing between different classes. A higher AUC value indicates better discrimination ability, with values closer to 1 indicating superior performance. By analyzing the ROC curves and AUC scores, we were able to assess the models' ability to accurately classify positive and negative instances, thereby informing decisions regarding model selection and deployment.

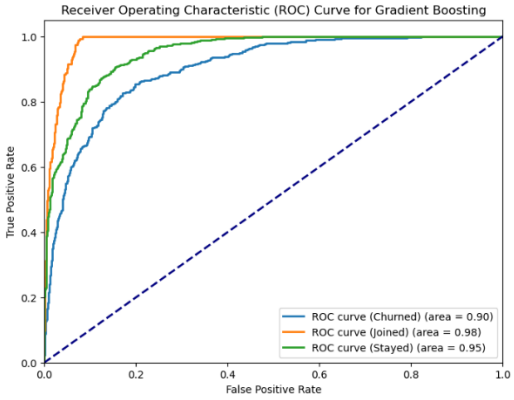

Figure 11: ROC curve for Gradient Boosting

A Gradient Boosting model's prediction ability across various consumer behavior's in the context of telecom churn is assessed using the ROC curves shown in Figure 17. The blue dotted curve, denoted as "ROC curve (Churned)" with an area under the curve (AUC) of 0.90, signifies the model's effectiveness in identifying churned customers, showcasing good discrimination ability. While not reaching excellence, this AUC value denotes a commendable capability in distinguishing churned from non-churned customers. Conversely, the orange curve labeled "ROC curve (Joined)" displays an AUC of 0.98, indicating the model's exceptional proficiency in discerning potential new customers from non-customers. This high AUC underscores the model's robust discrimination power in identifying customers likely to join or subscribe to new services. Lastly, the green curve, termed "ROC curve (Stayed)" with an AUC of 0.95, represents the model's aptitude in predicting customer retention, exhibiting very good discriminatory performance albeit slightly lower than the "Joined" model. These findings underscore the versatility and efficacy of the Gradient Boosting model in addressing diverse predictive tasks within the telecom churn domain, offering invaluable insights for operational decision-making and customer management strategies.

E. Feature Importance

To gain insights into the factors influencing customer churn, we analyzed the feature importance of the best-performing model. For the Gradient Boosting model, we visualized the feature importance using a bar chart, highlighting the relative importance of each predictor variable in influencing model predictions. Similarly, for the Logistic Regression model, we visualized the absolute coefficients of the features for each class, providing insights into the direction and magnitude of their impact on the predicted outcomes. By identifying the most influential predictors of churn, we gained valuable insights into the underlying drivers of customer behavior, enabling targeted interventions and strategic decision-making.

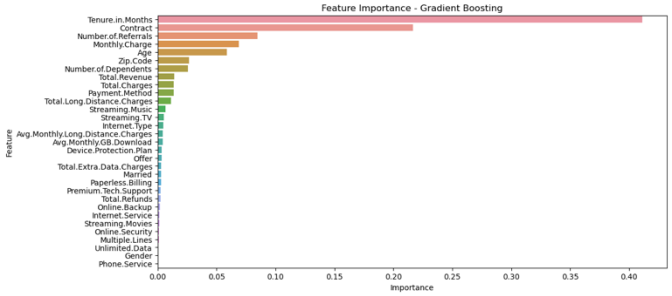

Figure 12: Feature Importance for Gradient Boosting

In our study, we conducted a comprehensive feature importance analysis on the Gradient Boosting model to uncover the primary drivers of customer churn within the telecommunications industry. The Figure 12 showcases the importance scores assigned to various features, providing insights into their relative influence on churn prediction. Remarkably, "Tenure in Months" emerges as the most critical predictor, with a substantial importance score of 0.411. This underscores the significant role played by customer tenure in determining churn likelihood, with longer-tenured customers exhibiting lower propensity to churn. Furthermore, "Contract" and "Number of Referrals" are identified as influential factors, highlighting the impact of contract terms and referral programs on customer retention efforts. Additionally, demographic attributes such as "Age" and "Marital Status" demonstrate notable importance, indicating the relevance of customer demographics in churn prediction models. Service-related variables such as "Monthly Charge" and "Internet Type" also exhibit considerable importance, emphasizing the influence of pricing and service quality on customer churn behavior. Moreover, the analysis unveils the significance of features related to billing preferences, technical support, and additional services in influencing churn decisions. This comprehensive examination of feature importance provides actionable insights for telecom providers, enabling them to prioritize retention strategies and tailor marketing efforts based on the most influential factors driving churn.

In evaluating interaction effects, on how payment method, contract type, monthly costs, and prior refund history collectively influence churn likelihood. The results underscored nuanced relationships between these factors and customer defection, revealing specific combinations that correlate with heightened churn propensity. Notably, customers on month-to-month contracts, experiencing elevated monthly costs, and with a history of refunds emerged as particularly vulnerable to churn, indicating the presence of synergistic effects amplifying risk within this segment. This insight highlights the necessity for targeted retention initiatives addressing multiple attributes simultaneously, tailored to mitigate churn risk effectively within these high-risk segments. By identifying and understanding the intricate interplay between various factors, telecom providers can devise proactive strategies to preemptively address churn drivers and bolster customer loyalty, thereby safeguarding revenue streams and fostering sustainable growth in a competitive market landscape.

VIII. Retention Model And Results

A. *Retention Model Development*

Retention strategy plays a pivotal role in mitigating customer churn, a critical concern for businesses across industries. In our study, we devised a comprehensive retention strategy model based on predictive analytics and feature importance analysis. Leveraging machine learning techniques, specifically Gradient Boosting, we predicted churn probabilities for individual customers using a rich dataset encompassing diverse demographic and behavioral attributes. Subsequently, we identified customers at high risk of churn, setting a threshold probability of 0.9 to define this segment. Our model then prioritized the top five features contributing to churn prediction, including tenure, contract type, number of referrals, monthly charge, and age. Each feature was associated with a distinct retention strategy tailored to its predictive significance. For instance, customers with shorter tenure received offers for loyalty discounts, while those with expired contracts were targeted with contract renewal incentives. Similarly, referral incentives, plan adjustments, and age-based promotions were offered to relevant customer segments. By aligning retention strategies with predictive insights, our model aims to proactively engage at-risk customers, enhance their loyalty, and ultimately drive business sustainability.

B. *Retention Model Evaluation*

The implementation of our retention strategy model yielded promising results in identifying and engaging customers at high risk of churn depicted in Figure 13. Through an exhaustive analysis of customer segments based on feature importance, we unveiled nuanced patterns and preferences driving churn behavior. Notably, our model's ability to tailor retention strategies to specific customer characteristics resulted in personalized interventions that resonate with individual needs and preferences. For instance, offering loyalty discounts to customers with shorter tenure acknowledges their nascent relationship with the brand, incentivizing continued engagement. Similarly, providing referral incentives capitalizes on existing social networks, fostering a community-driven approach to customer retention. Moreover, the scalability and adaptability of our model enable seamless integration into existing customer relationship management frameworks, empowering businesses to deploy targeted retention initiatives at scale. Overall, the application of our retention strategy model not only mitigates churn risk but also cultivates deeper customer relationships, laying the foundation for sustained growth and competitive advantage in dynamic market landscapes.

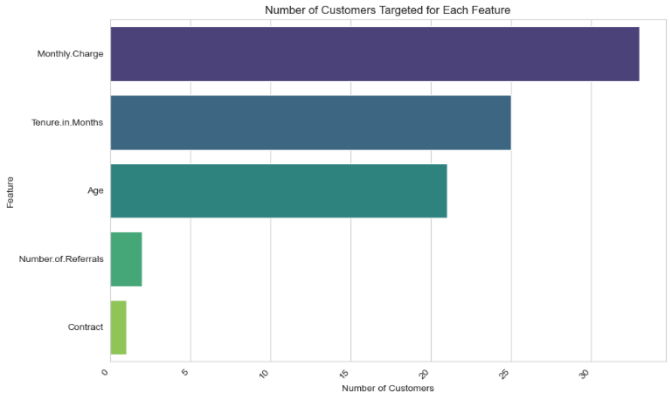

Figure 13: Number of Customers Targeted for Each Feature

IX. Limitations And Future Improvements

While the machine learning models demonstrated satisfactory performance in predicting the overall customer churn status, their effectiveness in accurately classifying the specific churn categories proved to be less robust. Despite employing rigorous hyperparameter tuning and optimization techniques, the models consistently struggled to differentiate between the various reasons cited for churn, including attitude, competitor influence, dissatisfaction, and others. The evaluation metrics, including accuracy, precision, recall, and F1-score, underscored the challenges faced by the models in accurately categorizing churn reasons, with performance metrics falling below desirable thresholds which is depicted in Table 2 and Figure 14.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	46.88%	38.83%	46.88%	36.31%
Random Forest	45.10%	30.95%	45.10%	33.25%
Gradient Boosting	44.39%	35.83%	44.39%	37.18%
SVM	47.77%	34.57%	47.77%	34.02%
k-Nearest Neighbors	39.39%	32.73%	39.39%	34.22%
Naive Bayes	33.69%	34.19%	33.69%	31.70%
Artificial Neural Network	40.11%	35.10%	40.11%	36.82%

Table 2: Performance Overview for 'Churn.Category'

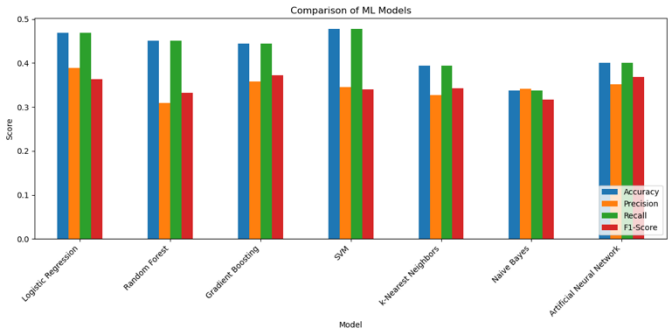

Figure 14: Comparison of ML models for 'Churn.Category'

One prominent factor contributing to the suboptimal performance of the models in predicting churn categories is the inherent class imbalance within the churn category data. With a significant portion of customers being classified under the 'N/A' category, the models encountered difficulties in discerning patterns and decision boundaries for the less represented churn categories. Addressing this imbalance through data augmentation techniques, such as oversampling minority classes or generating synthetic data, could potentially enhance the models' ability to learn discriminative features and improve predictive accuracy for all churn categories.

Additionally, the complexity and overlap in the underlying reasons for customer churn present another significant challenge for model performance. Customers may cite multiple or ambiguous reasons for discontinuing their service, making it challenging for the models to accurately classify churn categories based on distinct decision boundaries. Future research efforts could focus on incorporating advanced feature engineering techniques to extract more informative features related to customer behavior, preferences, and interactions, thereby enabling the models to better capture the nuanced reasons driving churn.

In pursuit of improving the models' performance in predicting churn categories, future work could explore ensemble methods and advanced modeling techniques, such as deep learning architectures and gradient boosting algorithms. Ensemble methods, including bagging and boosting, could harness the collective intelligence of multiple base models to improve predictive accuracy and robustness. Additionally, qualitative analysis techniques, such as sentiment analysis of customer feedback and reviews, could provide valuable insights into the underlying drivers of churn, complementing the predictive capabilities of machine learning models. By addressing these limitations and exploring new avenues for improvement, future research endeavors aim to enhance the models' ability to accurately predict churn categories, enabling more targeted and tailored retention strategies based on the specific reasons driving customer attrition.

X. Conclusions

This research paper has undertaken an in-depth analysis of customer churn within the telecommunications sector, applying advanced predictive modeling techniques to both anticipate and mitigate churn. The core findings reveal that the Gradient Boosting model, distinguished by its robust predictive accuracy of 85.05%, is particularly effective in identifying customers at risk of churn. Critical predictors such as tenure, contract type, and monthly charges have been identified as significant in influencing customer retention decisions.

The practical implications of these findings are substantial for telecom companies. By integrating the predictive models developed in this study, telecom companies can proactively identify at-risk customers and implement targeted retention strategies, potentially leading to significant cost savings and enhanced revenue through improved customer retention. This research not only supports telecom companies in their operational strategies but also contributes to the

academic discourse by refining the predictive modeling techniques used in customer churn analysis.

However, the study is not without its limitations. The challenges in predicting specific churn reasons suggest a need for more nuanced models that can handle complex, multi-faceted data inputs. Future research should explore the use of ensemble methods and deep learning to enhance predictive accuracy and robustness. Additionally, further investigation into the interplay of different customer features and their collective impact on churn could yield more comprehensive insights.

In conclusion, the issue of customer churn continues to pose a significant challenge in the competitive telecom industry. This research underscores the critical role of data-driven strategies and advanced analytics in addressing this challenge. By harnessing the power of machine learning and predictive modeling, telecom companies can not only anticipate customer behavior but also engage in more effective management practices to retain their customer base. Thus, this study contributes a significant step forward in the quest for sustainable business practices and enhanced customer relationship management in telecommunications.

XI. Research Website

https://mason.gmu.edu/~smantri4/

XII. References

[1] Bingquan Huang, Mohand Tahar Kechadi, Brian Buckley "Customer churn prediction in telecommunications" [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417411011353

[2] Ning Lu, Hua Lin, Jie Lu, Guangquan Zhang "A Customer Churn Prediction Model in Telecom Industry Using Boosting" [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6329952

[3] Samah Wael Fujo, Suresh Subramanian, Moaiad Ahmad Khder "Customer Churn Prediction in Telecommunication Industry Using Deep Learning" [Online]. Available: https://digitalcommons.aaru.edu.jo/cgi/viewcontent.cgi?article=1240&context=isl

[4] Utku Yabas, Hakki Candan Cankaya, Turker Ince "Customer Churn Prediction for Telecom Services" [Online]. Available: https://ieeexplore.ieee.org/document/6340176

[5] Kartekay Goyal, Kumar Kanishka, Kanisk Vasisth, Sahil Kansal, Ritesh Srivatsa "Telecom Customer Churn Prediction: A Survey" [Online]. Available: https://ieeexplore.ieee.org/document/9725621

[6] Pushkar Bhuse, Aaayushi Gandhi, Parth Meswani, Riya Muni, Neha Karte"Machine Learning Based Telecom-Customer Churn Prediction" [Online]. Available: https://ieeexplore.ieee.org/document/9315951

[7] Maw Maw, Su-Cheng Haw, Chin-Kuan Ho "Customer Churn Prediction in Telecommunication: An Analysis on Issues, Techniques and Future Trends" [Online]. Available: https://ieeexplore.ieee.org/document/9214592

[8] Yakub K.Saheed, Moshood A. Hambali "Customer Churn Prediction in Telecom Sector with Machine Learning and Information Gain Filter Feature Selection Algorithms." [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9655792

[9] Yashraj Bharambe, Pranav Deshmukh, Pranav Karanjawane, Diptesh Chaudhari, Dr. Nihar M. Ranjan "Churn Prediction in Telecommunication Industry Using Machine Learning Techniques" [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10080425

[10] Nur Idora Abdul Razak, Muhammad Hazim Wahid "Advanced Churn Prediction Models in Telecommunications Using Diverse Machine Learning Techniques" [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9642137

[11] Abhishek Gaur, Ratnesh Dubey "Predicting customer churn prediction in telecom sector using various machine learning techniques" [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8933783

[12] Ashish Sharmaa , Prafullit Shuklab, Mahendra Kumar Gourisariac, Bhisham Sharma "Telecom Churn Analysis using Machine Learning in Smart Cities" [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10085183