IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Research Article)

Check for updates

Predicting customer churn in subscription-based businesses using machine learning

Abdul-Waliyyu Bello [1, *], Idris Ajibade [1] and Darlington Ekweli [2]

[1] *Department of Mathematics and Statistics, Austin Peay State University, Tennessee, USA.*
[2] *Department of Healthcare Administration, University of the Potomac, Washington, USA.*

International Journal of Science and Research Archive, 2025, 16(03), 1024-1038

Publication history: Received on 16 August 2025; revised on 22 September 2025; accepted on 24 September 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.3.2664

Abstract

This study investigates the use of machine learning models for customer churn prediction in subscription businesses. The study investigates the influence of significant features such as tenure, monthly charges, and contract on churn prediction and how the two models classify churn and non-churn customers. The information contained considerable class imbalance with non-churn customers well outpacing churn customers, which proved difficult for accurate prediction. Despite this, Random Forest and Boost both exhibited strong classification performance with an accuracy rate of 79% and AUC of 0.83 and clearly showed that they were capable of identifying successfully the churn. The results show that Boost outperforms Random Forest with a slightly better recall and precision values, showing that it does better on the churn case detection without compromising with the balance offered with the non-churn customers. Analyzing feature importance, tenure, contract, and monthly charges appeared at the top rank among all predictors of churn, as suggested by previous literature based on customer interaction and monetary concerns. The study also found that customer support-related features, such as Tech Support, were critical to the churn prediction mechanism. The results suggest that companies are able to use machine learning models to identify which customers will churn and formulate targeted retention efforts. Additional studies would make these models more streamlined and analyze additional customer behavior data for even better churn prediction, eventually aiding customer retention programs in subscription companies.

Keywords: Customer Churn; Machine Learning; Random Forest; XG Boost; Subscription-Based Businesses; Tenure; Monthly Charges; Contract; Churn Prediction; Customer Retention; Predictive Modeling

1. Introduction

Subscription-based business has emerged as leading revenue engines in today's highly competitive digital economy. Streaming platforms and telecommunication companies run models that guarantee regular revenue generation, as customers are likely going to subscribe when their current subscription ends (Kolomiets, et al., 2021). However, over the years, there has been an increase and intense competition among companies operating on subscription-based models. This has made every player in this business susceptible to customer abandoning the service of one provider for another (Dahlen and Mauritzon, 2023). A customer discontinuing his subscription could emanate from different reasons, particularly when customer feels dissatisfied with the services being rendered. Losing a subscriber may feel insignificant but this could have unsatisfactory influence on the company's revenue, which could lead to poor performance. As asserted by Mahesh et al. (2017), sectors with high acquisition costs are usually hit severely when customer churn occurs. Zhang (2023) argued that an increase in churn by 5% can lead to substantial reduction in a subscription-based revenue by 90%.

According to Altieri (2024), customers are unpredictable as their expectation continues to change over time, which puts businesses under pressure to be conversant with these changes to avoid unexpected switch of customers to their

[*] Corresponding author: Abdul-waliyyu Bello

nearest competitor. Khare and Arora (2024) were of the opinion that many organizations are still stuck with adopting reactive strategies to be aware of the departure of a customer, which in most cases fail to capture the early signal that a subscriber is likely to switch. By implication, customer churn is not only related to satisfaction and subscription rates, as there is another complex behavioral phenomenon attached to it. Businesses that run on subscription model believe that a subscriber is expected to return after exhausting his current plan, making it a long-term relationship (Musunuri, 2024). This relationship depends on several factors including service quality, customer satisfaction, frequent price review, perception of the subscription, and other factors not within the business capacity (Prabhadevi, et al., 2023). All these have contributed to the difficulty in predicting customer churn, particularly with qualitative analysis.

It is worth noting that the drivers of churn are not the same for every industry and company operating on subscription model. The factors that may induce a customer to leave in digital content platforms may differ from what triggers subscribers to disengage in telecommunication industry (Zhang, 2023). The advent of machine learning offers a different perspective to identifying customers that are likely to leave a subscription. With numerous machine learning algorithms, which does not necessarily depend on linear relationship, allows for prediction of customer churn through learning from existing data (Theodoridis and Tsakiris, 2021). Models such as decision trees, random forests, gradient boosting, and neural networks can learn from customer behavior logs, usage patterns, transactional histories, and support interactions to anticipate churn with increasing precision (Pedograph, et al., 2022). An additional advantage of the machine learning algorithms is being able to improve over time with the ability to be re-trained as new information is added. Subscription-based businesses, through machine learning algorithm can make valuable decision from real-time data, which becomes a key difference for the organizations facing churning challenges (Sai Mahesh, et al., 2024).

Despite growing popularity, machine learning application to churn prediction is plagued by a series of practical and methodological challenges. Oversimplified datasets with sparse features that ignore domain-specific intricacies or customer lifecycle stages are employed in a lot of academic and commercial implementations (Aljifri, 2024). In addition to inherent challenges such as class imbalance, where churners are a small proportion of users, that may skew model performance and yield misleading accuracy metrics. Black-box models, while powerful, also amplify interpretability concerns (Musunuri, 2024). This means decision-makers are unable to trust and act on predictions. Such disconnects point to the importance of research that combines rigorous modeling with attention to business relevance, interpretability, and fairness, especially as machine learning is being embedded in automated CRM and marketing systems (Kolli, Varadharajan, Ajith, & Kumar, 2025). Good prediction models for churn must balance both accuracy and explainability and contextual specificity. Although complex models such as Boost or deep learning architecture can offer marginal improvements in predictiveness, they must be supported by feature importance plots, SHAP values, or decision rules to allow for human interpretation (Gasi & Linderoth, 2024).

This study explores the application of supervised machine learning techniques to predict customer churn in subscription-based businesses using real-world behavioral and transactional data. Knowing that losing customers is costly, and that reactive retention strategies can only help so much, you can see how having a predictive approach is a more proactive, data-driven approach. This study aims to identify hurdles that detect early signs of churn so businesses can take preventative action that limits customer disengagement.

2. Literature Review

2.1. Customer Churn

Customer churn is a key issue for a subscription-based business model, as long-term revenue growth goes hand in hand with long-term customer retention. Churn is usually defined as when a customer stops using your service or subscription during a given period. However, churn is more than a simple "yes" or "no" answer; it's a combination of customer behavior, perceived value, intensity of engagement, and alternatives available on the market (Overbrace et al., 2013). Churn can be thought of as either voluntary churn, where users actively decide to cancel their service, or involuntary churn, which may take the form of failed payments or an account that has become inactive (Shaaban et al., 2012).

The significance of churn concerns its relationship with Customer Lifetime Value (CLV) and the asymmetry in cost between acquisition and retention. Ahmad et al. (2019) asserted that acquiring a new customer costs the company between 5 and 7 times more than keeping an existing customer. Reichheld and Sasser (1990) laid the groundwork for establishing that even the most modest change in retention rates can have an outsized impact on profitability. In this way, churn has become a substitute measure for customer satisfaction, product market fit, and organizational operational sustainability. What makes churn especially peculiar to subscription products and services is the recurring revenue of the customer, where the loss of one customer affects not only the current income but also cash flows expected in the future. This creates a unique urgency to create early detection of early churn signals as it is more effective and least expensive to impact retention efforts earlier in the churn process (Burez and Van den Poel, 2009).

Many organizations actually miss identifying these churn signals because they rely on lagging indicators of churn such as cancellation forms or not-renewal notices. Churn is not explained by either demographic or transaction. As emphasized by Amin et al. (2017), logins, features, and support, has greater predictive capability. As such, churn is therefore emphasizing more behavioral, dynamic, and high-dimensional data, and this has increased the need for intelligent, data-driven models that can monitor users based on continually evolving trends over time.

2.2. Traditional Approaches to Churn Analysis

Prior to machine learning, churn analysis was static, typically founded on general statistical methods and rule-based segmentation. Logistic regression was among the most widely used methods due to its simplicity and interpretability but foremost because it provided a means to estimate the likelihood of churn as a function of observable traits (Nesli et al., 2006). As with other statistical models, logistic models allowed firms to make inferences about the marginal changes in the likelihood of customer churn as a function of tenure, contract type, or average revenue per user (ARPU). Embedded in statistical methodology, logistic models also assume some level of linearity and independence on the part of predictors making them an imperfect fit for the complex, multi-dimensional, nature of customer behavior found in contemporary subscription-based engagement settings (Bucking and Van den Poel, 2005).

Besides regression-based methods, heuristic rules were also applied on customer relationship management (CRM). They included pre-decided, static values, where the customer was considered to be potentially churning if they did not log in for the past 30 days or purchase anything in the past 30 or 60 days (Ngai et al., 2009). These approaches were operationally successful, but very inflexible and therefore labelled involved but low-involvement users and overlooked subtle changes in user behavior that preceded love disengagement. RFM (Recency, Frequency, Monetary value) models were another standard procedure applied in mass retail and telecom settings. As beneficial as segmenting customers by participation is to facilitating focused contact, RFM cannot provide us with a predictive role in that it is limited by withholding the influence of interaction or situational factors from its process. These traditional approaches were marred by problems of imbalanced datasets where churners were a minority of the sample and thus provided biased models that leaned towards predicting the majority class of non-churn (Casement and Van den Poel, 2008).

Traditional churn analysis approaches are also reactive in nature, where risk is inferred based on behavior following disengagement, typically in the form of a customer departing. Legacy models leverage point-in-time snapshots of data, without consideration for the time-based nature of customer engagement. This typically positions organizations to act too late, after the customer has disengaged or decided to leave. Further, legacy models will rarely allow for custom retention approaches, but instead will group all at-risk customers as one homogeneous segment, blind to the fact that two segments may be churning for completely different reasons. This absence of segmentation and personalization diminishes the effectiveness of retention campaigns, and has the potential to disengage the customer when the intervention is perceived as irrelevant or ill-timed (Alcaraz, 2018).

2.3. Machine Learning for Churn Prediction

Machine learning (ML) is quickly heading towards a new paradigm in churn prediction with its scalable, flexible, and data-driven solutions to existing techniques. Most traditional churn forecasting techniques assume the form of stiff connections between input variables. However, ML algorithms are best equipped to identifying non-linear relationships, feature interactions, and hidden behaviors between recurring classes of "data" (Verbeke et al., 2012). In subscription business, where customer behavior is a mix of various factors (frequency of usage, past activity, service quality, and web behavior), this is an enormous source of flexibility regarding how well it can learn to predict churn more precisely and contextually.

The majority of churn prediction frameworks utilize supervised learning, in the guise of classification algorithms. Classification algorithms are trained using labeled data, i.e., where the churn results are known, such classification algorithms can then be utilized to make predictions about probabilities of churn for new instances. Examples of classification algorithms include decision trees, random forests, and gradient boosting machines (e.g., Boost), and all bring certain strengths to the table. While decision trees are interpretable, they sometimes overfit data. Random forests provide a more generalization with many different trees. The ensemble of boosting algorithms uses many models, iteratively reducing error (Hassouna et al., 2015). Settings where the data is unstructured or measurable in a sequence have also been tackled by recent work but at interpretability cost (Sarker et al., 2020).

The use of machine learning to churn modeling is not only facilitated by machine learning's capability for processing high-dimensional and heterogeneous data, ranging from demographic characteristics and purchasing behavior to clickstreams, app history, and customer support history, but also by the many characteristics of churn to which machine learning can respond. Amin et al. (2017) indicated that including behavioral features like login frequency, payment delinquency, and time-on-the-platform greatly enhanced the accuracy of churn prediction using a data set from the telecom sector. Similarly, in the software-as-a-service (SaaS) configuration, combining subscription records and in-app activity logs allowed companies to detect disengagement many weeks ahead of contract completion (Zhou et al., 2021). The flexibility of machine learning therefore renders it a prime choice for any company that has user habits with enormous variety and changing often and where the customer touchpoints are less homogeneous.

An additional benefit of ML models is their feature importance analysis which can provide potential insights to decision-makers. The tree-based class of models, such as random forest and Boost, generate ranked lists of the features used to assess churn, which not only improves the transparency of the model but can also generate specific retention action plans. Changing the pricing for the price-sensitive segment or improving the onboarding process for the users that had little earlier engagement and lack activity in order to better inform a retention plan. Tools for advanced interpretability such as SHAP (Shapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations), helped improve the transparency of black-box models to improve their reliability in business settings (Lundberg and Lee, 2017).

However, there are challenges to consider in the use of machine learning for churn prediction. Class imbalance can be a challenge since, frequently, the number of churners is less than the number of retained or non-churners, and this class imbalance typically applies pressure to model training, leading to high levels of accuracy with poor recall on the targeted minority (churn) class. The class imbalance can be addressed with numerous different techniques for managing it while developing machine learning models, such as by using SMOTE (Synthetic Minority Over-sampling Technique), or using under sampling or weighted loss functions (Fernández et al., 2018). Similarly, data drift, user behavior changing over time, can also lower the model's effectiveness unless retraining occurs periodically. On a real-world implementation level, operationalization requires integration with customer relationship management (CRM) systems and with a real-time data pipeline, which adds to any complexity.

However, there are also concerns to keep in mind in machine learning use for churn prediction. Class imbalance may be a problem since, in many cases, the churners are less in number than retained or non-churners, and this class imbalance tends to strain model training with very high accuracy but mediocre recall on the minority (churn) class of interest. Class imbalance can be addressed in numerous ways to tackle it while making machine learning models, by using SMOTE (Synthetic Minority Over-sampling Technique) or by using under sampling or weighted loss functions (Fernández et al., 2018). Data drift, user behavior changing over time, can also degrade the model's performance unless regular retraining is performed. At a practical implementation level, operationalization must be blended with customer relationship management (CRM) systems and with a real-time data pipeline, which adds to any complexity.

2.4. Key Challenges in Machine Learning-Based Churn Prediction

Although machine learning is a fantastic method of predicting customer churn, it has several extremely critical problems to use in real life that are not discussed widely. Most important among them, is whether or not available data is 'quality' and 'granularity', which specific attributes we've tracked. Data related to customers tends to be split between: billing systems; customer service records; digital marketing assets; etc., which leaves gaps or discontinuities in data across divisions. It's clear why omitting information about the most significant customer events or retaining lost subscriptions would impact model learning, or potentially cause some bias. Information gathered to support a variety of operational processes may lack the kind of organization, structure or balance required to support efficient model training, and therefore much work will be necessary in order to carefully preprocess or normalize it in order to become useful for churn formation models.

Even though machine learning models might offer advanced prospects for modeling and conflicting presumptions regarding non-linearities of data representation, they always heavily depend on relevant, expressive constructed features. Customer behavior could drastically change over time in subscription cases because of different currency rates, subscriptions or UI changes, seasons, etc., and therefore require time-based features that also fluctuated over the customers' lifecycle. Common static based attributes like age, gender or tenure do not necessarily represent good inferential decision points. It is still an art blend of science and to design time based or frequency based, or session-based attributes that do-good behavior change capture, and low representation of attributes usually leads to massive amounts of underfitting, despite the sophistication of the algorithms.

Churn is not necessarily going to follow predictable patterns, especially in freemium or hybrid subscription models. A free user lowering usage may not be developing risk of churn, while the same trend in a paid user can signal cancellation to occur soon. The lack of shared churn signatures in various tiers of a subscription (or product sets) adds to the friction potholes of generalizing models. A model trained for high-end subscribers will not generalize properly if applied to low-end users and requires either model tailor-made to the segment/subscription or might require utilization of hierarchical model structures.

The transferability of churn models across industries is another area where deficiencies differ. A model built for a telecom company may not be applicable to a SaaS business because of differences in "usage" signals, customer engagement expectation sensitivity, and different drivers of churn. The problem is that although churn can occur to the same firm, models constructed using old data are not going to function when business must change from a price consideration or user interface frameworks. This does not make models reusable and offers a demand for repeated retraining and retention, leading to increased operational expense and technical debt.

Making operational use of churn models to business processes becomes a real bottleneck to organizations. Even the most excellent churn model with highest predictive performance relies on timely, actionable insights to be able to provide value. Timely and automated decisions to push predictive outputs, start a marketing campaign, create automated offers, or even initiate a customer service call do not happen on the fly. Since they usually do not allow these live loops, nor even have such an ability by default, organizations introduce delays between when the prediction is made and when the prediction is acted upon. They can easily become obsolete or stale without strategic partnerships between data science units and business units for utilization of timely actionable and predictive outputs.

2.5. Theoretical Review

The Customer Lifecycle Theory suggests that customer relations are at five stages: acquisition, onboarding, active use, potential decline, and renewal or churn (Kotler and Keller, 2012). Each stage of the customer lifecycle demonstrates some behaviors that are indicators of the different levels of relationship, and each of these can be operationalized as a feature that can be encoded in a predictive model framework. Churn across various stages of relationships is anticipated at the transition from 'active' to 'at risk', when interaction has just started to decline into the ill path of the current relationship. Machine learning models, particularly those utilizing time-series or sequential form of features, are able to identify a mapping of the lifecycle development and ascertain initial warning signs of disengagement. However, it is important to understand the theory behind the stages of the lifecycle to recognize which behaviors are critical, which feature is notable, and when to take the action.

Combined with this is Relationship Marketing Theory, in which customers' engagement over simple transacting is more prioritized (Morgan and Hunt, 1994). In this case, churn is a matter of trust, satisfaction, or value perceived decreasing. The theory that emerges shows that service quality, emotional connection, and complaint responsiveness, are monumental variables in retention. Though this is anecdotally supported, these would be used to estimate for, in the building of machine learning models. The subtlety can percolate through in proxies e.g., rate of support center contact, net promoter score (NPS) or possibly motility a longitudinal customer churn after complaint follow-up. Furthermore, relationship marketing theory has built in an acceptance that churn is not equal. Some churn is caused by dissatisfaction with the service, but some churn is merely caused by situational factors. This distinction calls for a new focus on creating interpretable models that not only predict churn, but explain it.

In terms of data science, utility theory and cost-sensitive learning are useful concepts. Utility theory operates on the premise that a customer primarily judges an existing subscription based on estimated future worth against the cost. If a customer receives the impression of utility less than an individual threshold, withdrawal or churn will result. This is part of the role of usage measures of some kind e.g., lower time-on-platform or decreased platform interaction on features that presumably take place prior to cancellation. Cost-sensitive learning understands that incorrectly labeling a churner is many orders of magnitude more important than incorrectly labeling a non-churner.

2.6. Empirical Review

Numerous empirical researches have been conducted to evaluate machine learning-based models in customer churn prediction across various industries, such as telecommunication, banking, and Software as a Service platform. The IBM Telco Customer Churn dataset is one of the best-known datasets, where numerous classification models have been compared. Idris et al. (2019), contrasted Support Vector Machines, Random Forest, and Logistic Regression on the IBM Telco Customer Churn dataset and found that ensemble models are more stable and accurate at predicting categorical customer behavior. The said studies did include interpretability analysis, thereby discouraging practical application in real business settings.

In the same way, Amin et al. (2017) used decision tree-based models on a telecom churn dataset and found contract type, service failures, and customer tenure to be the strongest predictors for churn. The findings also pointed out the importance of balanced datasets and preprocessing, but the authors did not investigate temporal features nor conducted lifecycle segmentations. Similarly, Sarker et al. (2020) used deep learning methods and reported high precision and recall rates, but their deep learning approach relied on complex architectures and extensive computing resources prompting concerns about the viability of smaller to medium businesses with fewer infrastructure resources.

Within the SaaS industry, Zhao and Wang (2021) used gradient boosting machines in their analysis of software user churn, where they investigated engagement signals, i.e., regular logging in, completion of tasks within onboarding, earlier in the customer lifecycle that were highly predictive of churn. However, although their work was high in predictive validity, the authors only used their own proprietary data which limited the potential for reproducibility. Zhang et al. (2022) that utilized feedback from clickstream data and navigation data within apps, achieved predictions days or weeks in advance of churn. Nevertheless, while the models were excellent in terms of accuracy, there was also a level of criticality presented by the authors that concern for fairness was incomplete because they did not consider beyond the mechanisms of data representation and towards the potential level of bias against low or sparse engagement.

2.7. Gaps in the Reviewed Literature

While use of machine learning for churn prediction is growing, there is still a rather big body of literature based on more general models that do not take into account the advanced nature of subscription-type products. In fact, the majority of the papers provide empirical studies that examine static databases, such as those of telecommunication or banking industries, without addressing time-dependent behavior factors which delineate customer lifecycle changes. This constrains the model's ability to detect early warning signs of a churn event, and therefore perform proactive interventions. Besides, although the majority of research papers do involve ensemble models such as Random Forest and XG Boost, due to their intrinsic prediction quality, few utilize an interpretation model that informs the business user, for what reasons a customer might churn, I would argue that such an explanation is lost in not having it, and diminishes the practical business utility of these models. Without the understanding of reason behind decision-making, the company will not be able to translate the predictions into any policy adjustment.

Both ethical and operational aspects of churn prediction remain unmapped. Most studies prioritize accuracy over any potential bias in the model prediction, particularly for rare users, low-income, or minorities. The discussion is also limited on the challenge of deploying ML churn models into live business systems, such as real-time data streamed to the model, how to push predictions back into the system, and how to retrain the model on finite events due to behavior drift. Similarly, industry-specific churn drivers, such as freemium to paid consumers in SaaS, or billing cycle mismatch in media subscription are rarely considered by churn modeling. It becomes necessary to fill these gaps as an addition to existing body of knowledge.

3. Methodology

3.1. Research Design

This is predictive and quantitative study design, using machine learning techniques to predict customer churn for subscription businesses. A data-driven study attempts to discover the customer behavior and customer demographics data in order to develop a predictive model (Khare and Arora, 2024). The research adopts the supervised learning approach based on past data and the resultant churn observations that apply machine learning to build and therefore predict future churn. The models will include Random Forest, Boost, and others, and are contrasted based on how well they can predict customer churn. The research design provides a systematic, organized, and scalable method to gain insights and reduce churn in subscription schemes (Mohan & Jadhav, 2022).

3.2. Data Sources and Collection

This study made use of publicly available data from a dataset posted on Kaggle called Telco Customer Churn. It contains details on 7,043 customers of a single company and knows 21 features such as demographics of the customer, what products and services they were subscribed to, how they utilized the services, and if they churned or not. The features comprise both numerical distribution variables and categorical distribution variables that give a complete description of what is explanatory in order to predict churn by a customer.

This dataset is enormously valuable as it contains a mix of customer behavior and churn drivers of many customers which is typical of usual customer practice, it's representative of usual practices to churn prediction in a subscription business. The extent of detail and popularity for public utilization makes it eligible for appropriate analysis resulting in predicting purchasing behaviors (Bhatt & Nagvadia, 2021).

3.3. Data Preprocessing

The data was preprocessed to clean it for machine learning. Missing values were treated with imputation techniques or row removal where possible (Alam, 2023). Categorical features like contract type, payment type, and gender were converted to machine-readable format by one-hot encoding them. Numeric features like monthly charges and total charges were normalized to have the same scale applied to all predictors. Outliers were identified and processed so that the model would not be biased (Dash, Behera, Dehuri, & Ghosh, 2023). The ultimate result was a dataset of 7,043 records and 20 features to train the model, which was created after pre-processing.

3.4. Feature Selection

The study performed feature selection to render the most important predictors of customer churn as much as possible. This began with all of the features, but cautious were taken to remove duplicate features that were highly correlated (Shrestha, 2020). This study tested if features with very high multicollinearity can be removed according to a correlation matrix and only kept those features with greater predictive power. RFE and Random Forest feature importance were attempted to see if they will enhance the feature set (Michelucci, 2024). The final selected features to utilize tenant, service usage, payment method, and contract type, which remain the strongest predictors of churn.

3.5. Machine Learning Models

This study entails using three machine learning models for predicting customer churn: Logistic Regression as the baseline model, Random Forest Regressor and XG Boost Regressor. Logistic Regression has been added since it presented a simpler and easier-to-explain model to act as a baseline for comparison (Aljifri, 2024). Random Forest Regressor and XG Boost were employed as ensemble algorithms to address the complexity of the data structure and non-linear interactions among the multiple features and their intersections (Michelucci, 2024). The models would be trained using the training data dataset (80%) and tested on the test dataset (20%). Research was used for the hyperparameter optimization, and model performance measures were considered.

3.6. Model Evaluation

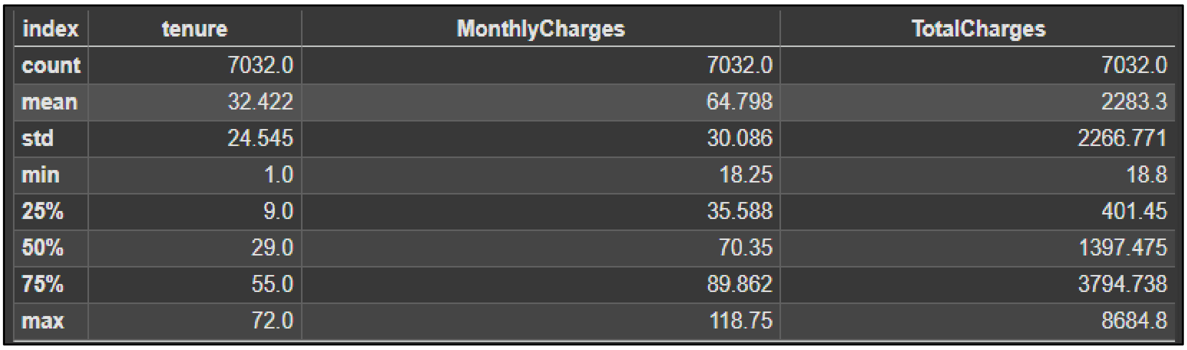
Model performance was evaluated using a set of performance metrics that would measure performance in models for predicting churn. Precision and recall metrics were utilized to measure the ability of the model to classify correctly the churners and not classify incorrectly the non-churners as churners, respectively (Gasi & Linderoth, 2024). Accuracy was utilized to measure the ratio of correct classifications to gauge how much the model was precise. An F1-score (harmonic mean of precision and recall) measure was used as precision may be good at certain times but recall is bad, and F1 would consider the trade-off (Zhang, 2023). AUC - ROC was used to illustrate trade-offs of true positive rate and false positive rate, and assess characteristics of discriminative ability. The metrics were important to be able to determine the strengths and weaknesses of a customer churn classification model and how accurate the overall prediction will be.

3.7. Ethical Consideration

Ethics was a major concern throughout the course of the study. This was most applicable to model explainability and data privacy. The dataset in use, used for this study, was openly available and contained no personally identifiable information (PII) and, therefore, met contracting, law, and data collection acts. The initial dataset adhered to the procedural process of gathering informed consent for data collection and applying anonymization methods. The study also bartered general principles of fairness by attempting to eliminate potential biases found in the data set like, imbalance of class issues, bias against minoritized and/or underrepresented groups. Also, prioritizing interpretability of machine learning models while making forecasts so that decision-makers could reasonably understand, interpret, and react to forecasts, thus allowing a responsible business practice capturing the use of the model.

4. Results

This section employs exploratory data analysis to inspect and describe the features before proceeding to adopting the adopted machine learning techniques for churn prediction.



index	tenure	MonthlyCharges	TotalCharges
count	7032.0	7032.0	7032.0
mean	32.422	64.798	2283.3
std	24.545	30.086	2266.771
min	1.0	18.25	18.8
25%	9.0	35.588	401.45
50%	29.0	70.35	1397.475
75%	55.0	89.862	3794.738
max	72.0	118.75	8684.8

Figure 1 Descriptive Analysis

Descriptive statistics for tenure, Monthly Charges, and Total Charges are displayed, indicating a mean tenure of 32.42 months with a standard deviation of 24.55, suggesting relatively moderate variability around that mean. The mean Monthly Charges is 64.80 with a standard deviation of 30.09, indicating there is fairly large variability in the Monthly Charges of the customers. The mean Total Charges is 2283.3 with and standard deviation of 2266.77, indicating that there is extremely large variability in the Total Charges. These values indicate varying customer behaviors in terms of length of subscription, monthly charges per customer, and total amount paid by customers.

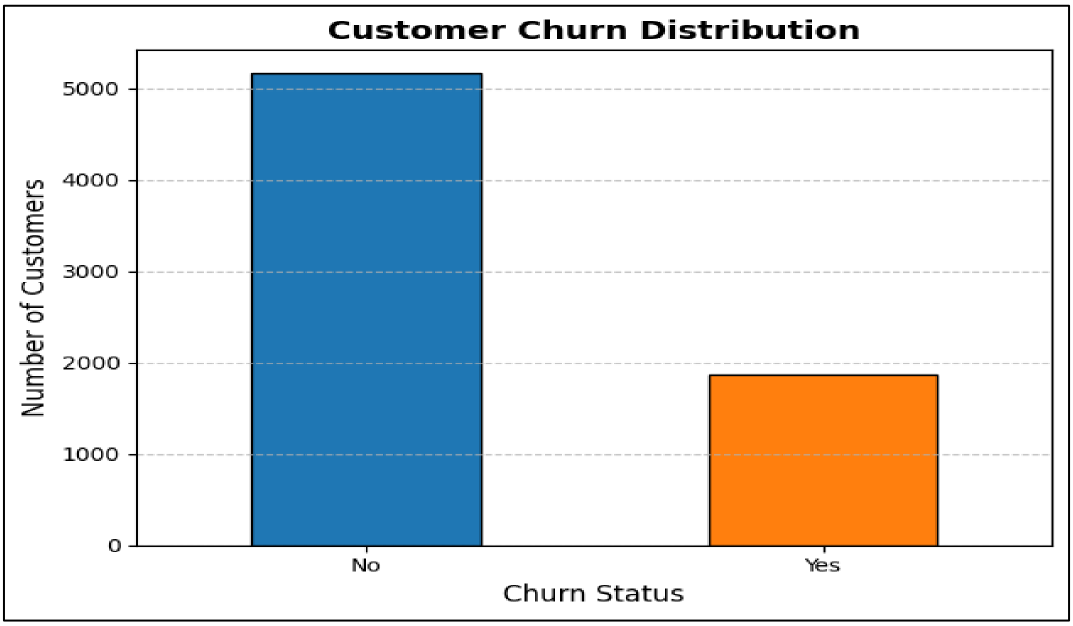


Figure 2 Distribution of Customer Churn

The bar chart illustrates a highly imbalanced distribution of customer churn. The majority of customers (approximately 5000) fall under the "No" churn status, indicating non-churn customers, while only a small portion (around 1000) are classified as "Yes" for churn.

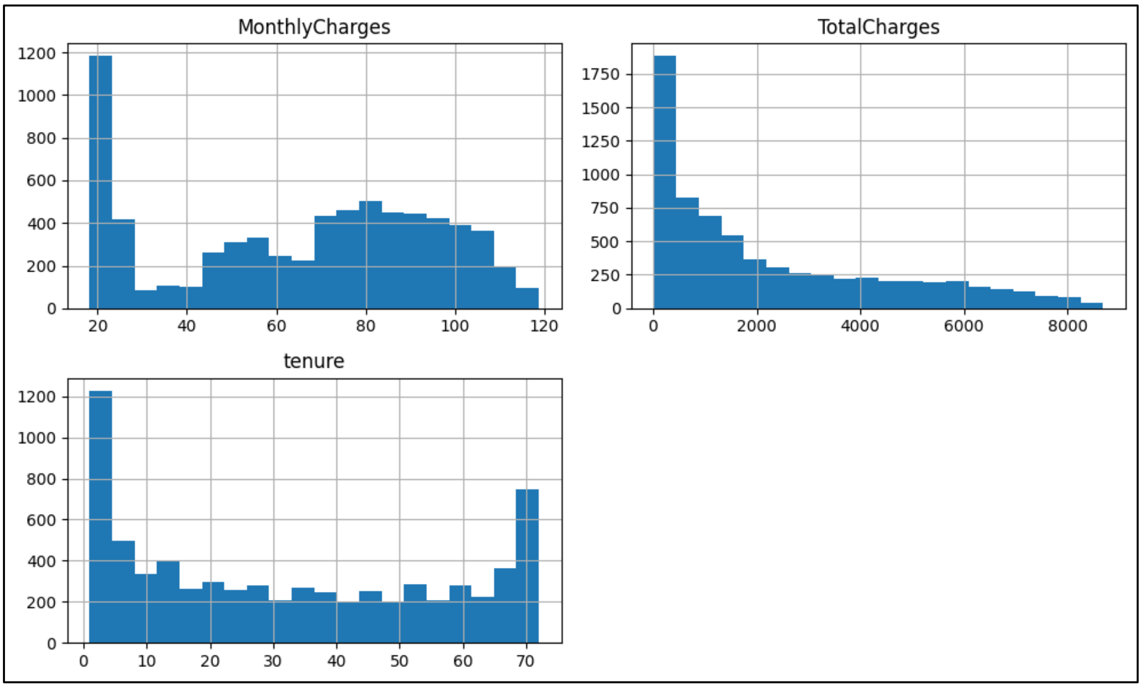


Figure 3 Histograms

Figure 3 displays the histograms for Monthly Charges, Total Charges, and tenure, showing notable distributions. Monthly Charges are predominantly clustered around 20, with a few customers having higher charges, while Total Charges exhibit a right-skewed distribution, with most customers having lower charges and a few with significantly higher totals. The tenure histogram reveals a bimodal distribution, with large groups of customers having either very short or long tenures, and fewer in the middle. These distributions suggest considerable variability across customers, which may influence the modeling process and interpretation of churn patterns.

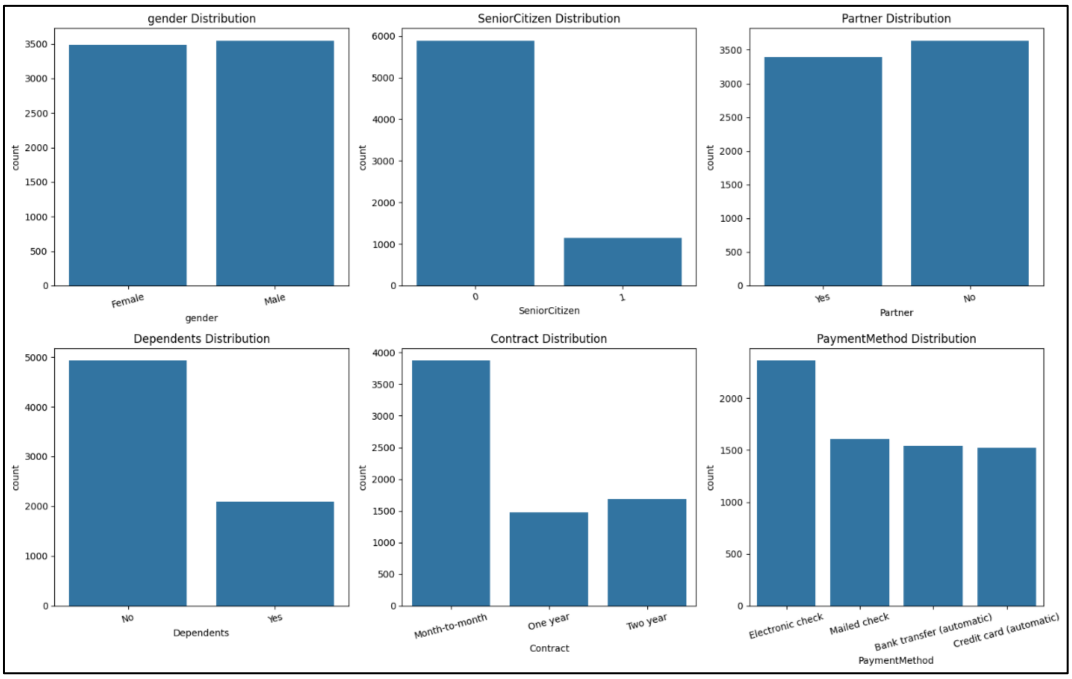


Figure 4 Distribution of Categorical Features

Figure 4 presents the distributions for customer demographics and contract or payment characteristics. The gender distribution highlights near balance with slight variance for males and females. Most customers also are not senior citizens, with reasonably large proportions of customers having partners versus not having partners. The Dependents distribution indicates that most customers do not have dependents. The Contract types are mostly month to month and fewer are on one-year or two-year contracts. The Pay Method distribution indicates a strong preference for electronic checks, i.e., with a non-negligible amount of customers using other methods, such as mailed checks and credit cards. These characteristics could contribute potentially to churn patterns.

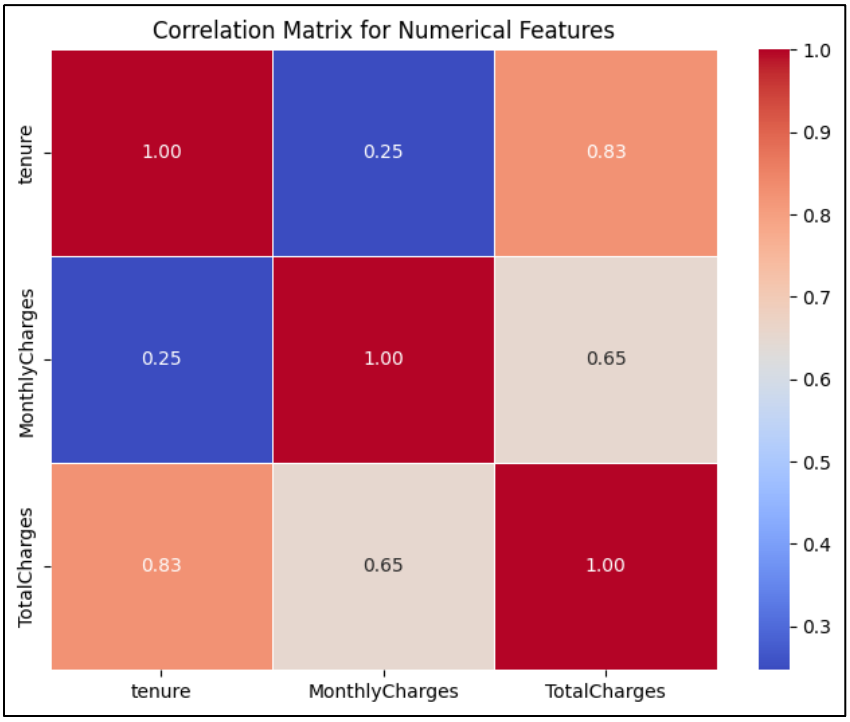


Figure 5 Correlation Matrix for the Numerical Features

The correlation matrix in Figure 5 displays the relationships between the numerical features: tenure, Monthly Charges, and Total Charges. There is a strong positive correlation between tenure and Total Charges (0.83). This suggests that customers who have a longer tenure also tend to have higher total charges. There is likewise a moderate correlation between Monthly Charges and Total Charges (0.65). This means that higher monthly charges are associated with higher total charges though, again not to the same extent as tenure. The relatively weak correlation between tenure and Monthly Charges (0.25) indicates that the length of customer tenure does not heavily influence the monthly charges. Given the high correlation, monthly charges were retained as part of the features, while total charges were dropped to avoid problem of severe multicollinearity.

4.1. Model Performance Evaluation

4.1.1. Hyperparameters

Table 1 Best Hyperparameters

Hyperparameters	Random Forest	XG Boost
max depth	10	3
max features	sqrt	-
min samples leaf	2	-
min samples split	2	-
n estimators	200	100
Col sample by tree	-	0.8
learning rate	-	0.1
subsample	-	0.8

The study employed hyperparameters in order to minimize overfitting while optimizing model performance. The fitted model identified as the best hyperparameters

As reported in Table 1, the best hyperparameters for the Random Forest model suggest a focus on deeper trees (adept = 10), more trees (estimators = 200), and a balanced feature selection (max_features = 'sqrt'). In contrast, XG Boost prioritizes shallower trees (adept = 3), a smaller learning rate (0.1), and enhanced generalization through feature subsampling (subsample = 0.8) and column subsampling (colsample_bytree = 0.8).

4.1.2. Evaluation Metrics

These results present the performance metrics for the machine learning models adopted in this study. More specifically, it shows how the baseline models, which logistics regression and the fine-tuned model in Random Forest and XG Boost performed with respect to the prediction of customer churn.

Table 2 Evaluation Metrics

Metric	Logistic Regression	Random Forest (Tuned)	XG Boost (Tuned)
Accuracy	0.79	0.79	0.79
Precision	0.63	0.64	0.64
Recall	0.50	0.48	0.49
F1-Score	0.56	0.55	0.55

All three models were evaluated on accuracy, precision, recall and F1-score. All three models have the same accuracy 0.79, meaning that on average their performances are equal. But the baseline model Logistic Regression scored precision of 0.63 and recall of 0.50, resulting in F1-score of 0.56. Random Forest (Tuned) precision was 0.64 and recall was 0.48 and XG Boost (Tuned) precision was 0.64 and recall was 0.49, both with approximate F1-scores of 0.55. Logistic Regression model results are a bit higher than Random Forest (Tuned) and XG Boost (Tuned) but falling metric after tuning or hyper-parameters rules repeat an actual value obtained from each evaluate method. These results suggest that the models are similar in nature and hyper-parameter tuning has diminishing returns in some of the metrics.

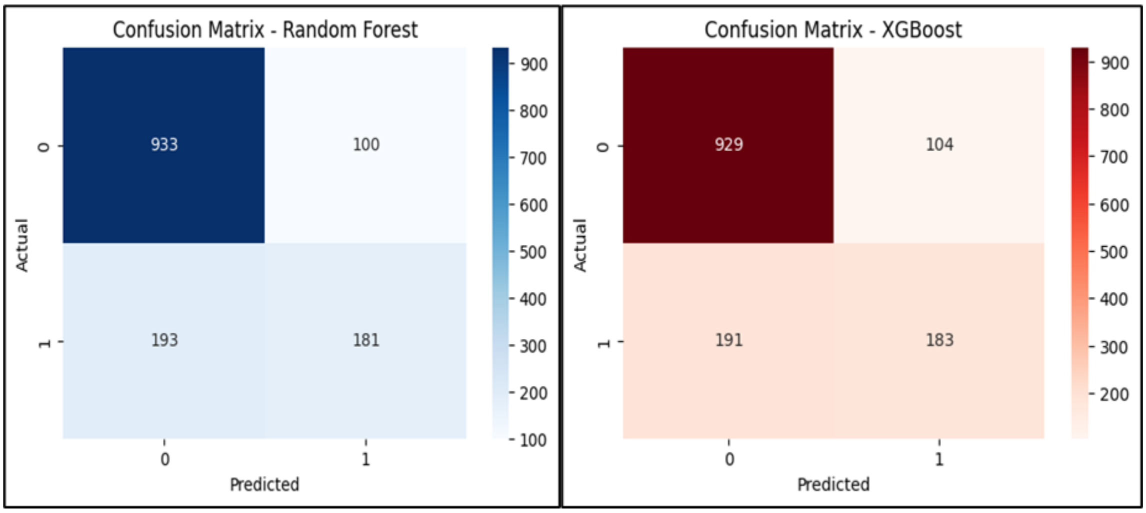


Figure 6 Confusion Matrix

Random Forest confusion matrix reveals an acceptable level of performance in predicting both the classes. The model accurately predicted 933 non-churn instances as true negatives, while it also correctly predicted 100 churn customers as false negatives through misclassifying them as non-churn. The modeling algorithm also accurately predicted 181 churn instances (true positives,) but also inaccuracy in predicting 193 churn instances as non-churn (false positives.) The number of non-churn instances suggests that the modeling algorithm is perhaps performing better with non-churn instances but misclassifying true churn instances.

The XG Boost modeling algorithm identified 929 as non-churn, incorrectly labeling 104 churn customers as non-churn. In its associated churn case prediction, the model algorithm accurately predicted 183 churn and incorrectly predicted 191 churn cases as non-churn. XG Boost performed similarly to Random Forest, with slightly more true positives suggesting slightly improved churn accuracy. Both models appeared not to be making adequate predictions regarding churn outcomes and may need to be upgraded for better churn prediction recall.

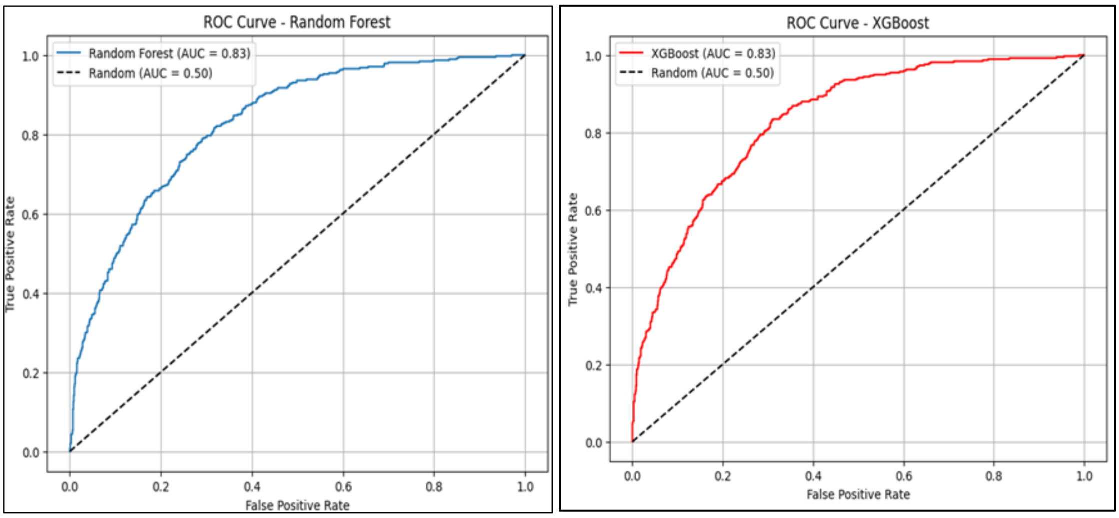


Figure 7 ROC Curves

As presented in Figure 7, Random Forest and XG Boost both exhibit strong classification performance based on the ROC curve results. Random Forest classifier performed at AUC 0.83, much higher than baseline random classifier (AUC = 0.50), since the curve sharply rises in the left column of the chart. XG Boost performed at AUC 0.83, reporting similar performance to Random Forest. Both models created a relatively high rate of true positives with few false positives and good discrimination between churn and non-churn cases, indicating both models' good predictive capability in customer churn prediction.

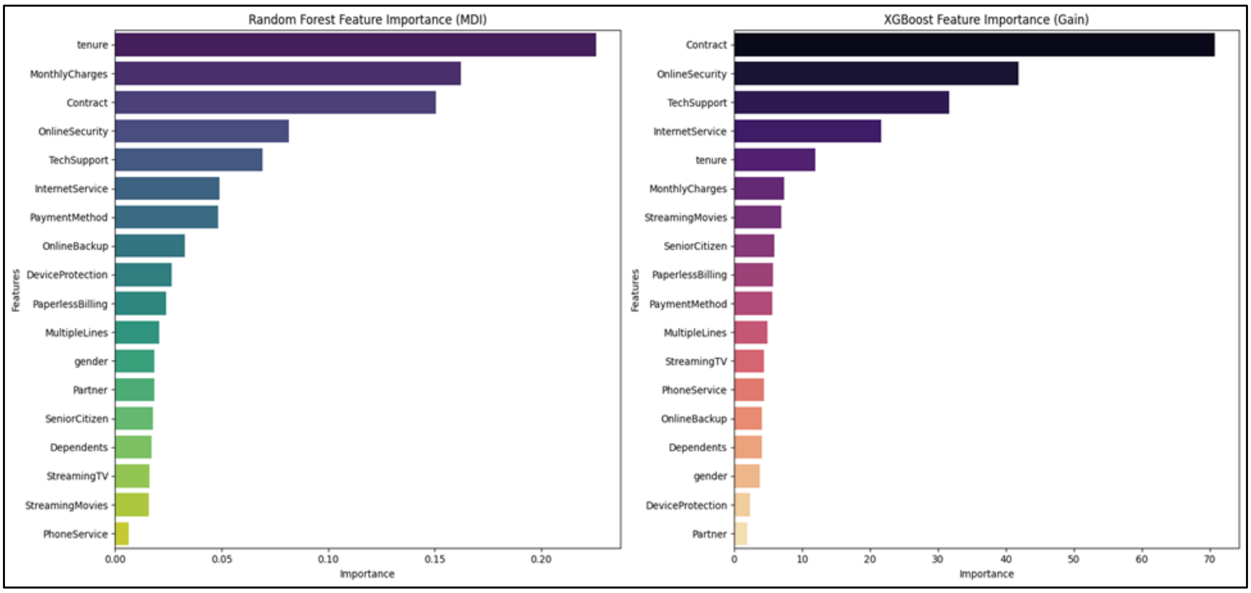


Figure 8 Feature Importance

Random Forest and XG Boost feature importance scores have different results for which were the important variables in predicting customer churn. For Random Forest, the most important features were tenure, Monthly Charges, and Contract. Tenure had the highest feature importance score, with Online Security, Tech Support, Internet Service also being important features in predicting customer churn. On the other hand, XG Boost placed a heavy degree of importance on Contract, Online Security, and Tech Support, with reduced importance for tenure. The implication here is that, while both of the two models consider elements of customer engagement (i.e. tenure and support services), Random Forest places more importance on contract type and charges, while XG Boost places heavy emphasis on the contract and security components.

5. Discussion of Findings

This research evaluated the applicability of machine learning models, specifically Random Forest and XG Boost, within a subscription company to forecast customer churn. A complete analysis yielded a number of interesting findings regarding model performance and the main predictors of customer churn. For instance, the data demonstrated that class distribution was highly imbalanced in favor of non-churn customers over churn cases, as others also experienced a similar class imbalance when trying to perform a similar customer churn prediction competition task (Bhatt & Nagvadia, 2021).

Descriptive statistics confirmed tenure, Monthly Charges, and Total Charges were predictors used to differentiate between churn and non-churn customers. The similarity between tenure and Total Charges variable with strong correlation confirms that a customer with longer tenure also had larger total charges. The distribution chart confirmed the class imbalance discovery and confirmed the challenge in understanding churn customers in a non-churn customer dominated dataset. Overall, Random Forest and XG Boost performed with excellent accuracy despite the imbalanced dataset, and their models were able to reach an AUC of 0.83, thus confirming that their models could distinguish between churn customers and non-churn customers.

Performance measurements showed clear differences in model performance. XG Boost was slightly more precise and had slightly higher recall than Random Forest, which indicated that XG Boost was better at classifying churn. Both models performed equally well at 79% accuracy, which showed outstanding overall performance. Confusion matrices reported that both models did a good job in classifying non-churn customers, but XG Boost did a bit better job in minimizing False Positives and False Negatives. This is particularly useful when the data is under affliction of class imbalance.

Feature importance analyses showed the top churn predictors to be Contract, tenure, and Monthly Charges. The predictors confirm other studies in the literature that identified customer activity, as well as financial metrics, to help predict churn (Kolli, Varadharajan, Ajith, & Kumar, 2025). Other predictors such as Partner and gender had low effects, which implies they were not as crucial to the analysis. These findings make predicting customer churn more understandable and provide even stronger evidence that the solution to predicting and preventing churn in subscription-based companies lies in machine learning.

6. Conclusion

This study demonstrated the ability of machine learning models, Random Forest and XG Boost, to predict customer churn in subscription businesses. The algorithms both had high performance metrics of 79% accuracy and an AUC of 0.83, demonstrating their ability to distinguish between churn and non-churn customers well even after they identified class imbalance. Key features such as tenure, Monthly Charges, and Contract proved to be robust predictors, as would be expected from standard literature on customer interaction and financial aspects being keys to churn forecasting. XG Boost showed a slight edge over Random Forest in terms of precision and recall, indicating its greater ability to handle the class imbalance. The study results emphasize the potential of machine learning models to improve the precision of churn prediction and inform retention business strategies.

Given the insights obtained from the study, the following are the recommendations

- Organizations need to keenly study how they can find additional ways to further improve engagement, as it is biased towards the available resources with lower tenure customers, given that tenure was the remarkable dominant predictor of churn. Some type of personalized message systems; and smart reminders have the potential to retain consumers engaged with something to resort to and strengthen their interaction with such a company that can prevent churn.
- With Monthly Charges as important predictors, it would be desirable for companies to consider their price policy, to see whether the customer feels rewarded appropriately for having the service. There needs to be differentiation in reward to be a long-term customer.
- Since attributes like Contract and Tech Support were the largest in support of predicting churn, companies must ensure that the customer was sufficiently supported and had an easy access to timely solutions to

problems. A 24/7 system of support with a better-quality service faster and more reliable might provide a customer with peace of mind - this would help retain them.

- Finally, firms need to determine what is characteristics would create chances that the customer is at high risk of churn. Firms need to be capable of seeing what characteristics would create churn and react. Preventive moving policies, like comprehensive retention offers or calls, could help de-escalate churn for customers that are at risk of churn.

Compliance with ethical standards

Disclosure of conflict of interest

The authors confirm that there is no conflict of interest to be disclosed.

References

[1] Alam, S. (2023). An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. Decision Analytics Journal, 9(1), 100341.

[2] Aljifri, A. (2024). Predicting Customer Churn in a Subscription-Based E-Commerce Platform Using Machine Learning Techniques. Dalarna Research. Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:du-48495

[3] Bhatt, V., and Nagvadia, J. (2021). Factors Influencing Consumer's Online Buying Behavour: An Empirical Study. Journal of Management, 16-65.

[4] Dahlen, D., and Mauritzon, W. (2023). Machine Learning-based Prediction of Customer Churn in SaaS. Journal of Computer Science, 48(1), 1-86.

[5] Dash, C. S., Behera, A. K., Dehuri, S., and Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. Decision Analytics Journal, 6(2), 100164.

[6] Gasi, I., and Linderoth, P. (2024). Combining ML Models for Churn Prediction and Customer Retention Strategies within Digital Subscription Services. Technology, School of Electrical Engineering and Computer Science, 3(7), 1-16.

[7] Khare, P., and Arora, S. (2024). Predicting Customer Churn in SaaS Products using Machine Learning. International Research Journal of Engineering and Technology, 11(5), 754-765.

[8] Kolli, M., Varadharajan, N., Ajith, K., and Kumar, K. D. (2025). Customer Churn Prediction in Subscription-Based Services. International Conference on Recent Trends in Machine Learning, 1(1), 247-257.

[9] Kolomiets, A., Mezentseva, O., and Kolesnikova, K. (2021). Customer Churn Prediction in the Software by Subscription models IT business using machine learning methods`. Information Technologies: Theoretical and Applied Problem, 4(14), 1-10.

[10] Mahesh, B. S., Jagadeesh, B., Gowtham, A., S, R. C., Kumar, K. K., and Kishore, R. S. (2017). Predicting Customer Churn in Subscription-Based Enterprises Using Machine Learning. Evolutionary Artificial Intelligence, 26(1), 365-377. doi:https://doi.org/10.1007/978-981-99-8438-1_26

[11] Michelucci, U. (2024). Fundamental Mathematical Concepts for Machine Learning in Science: Feature Importance and Selection. Springer, 1(1), 229-242.

[12] Mohan, M., and Jadhav, A. (2022). Predicting Customer Churn on OTT Platforms: Customers with Subscription of Multiple Services Providers. Journal of Information and Organization Sciences, 46(2), 433-451.

[13] Musunuri, A. (2024). Machine Learning Model for Predicting Customer Churn in Subscription Based Business. International Journal of Artificial Intelligence and Machine Learning, 3(2), 211-220.

[14] Peddarapu, R. K., Ameena, S., Yashaswini, S., Shreshta, N., and PurnaSahithi, M. (2022). Customer Churn Prediction using Machine Learning,. 2022 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India,, 1035-1040. doi:10.1109/ICECA55336.2022.10009093.

[15] Prabadevi, B., Shalini, R., and Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. International Journal of Intelligent Networkds, 4(1), 145-154.

[16] Sai Mahesh, B., Jagadesh, B., Gowtham, A., Seshagiri Rao, C., Kumar, K. K., and Kishore, R. S. (2024). Predicting Customer Churn in Subscription-Based Enterprises Using Machine Learning. Cryptology and Network Security with Machine Learning,, 3(13), 365-377.

[17] Shrestha, N. (2020). Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics , 8(2), 39-42.

[18] Theodoridis, G., and Tsadiras, A. (2021). Using Machine Learning Methods to Predict Subscriber Churn of a Web-Based Drug Information Platform. Open Science, 3(1), 581-593.

[19] Zhang, B. (2023). Customer Churn in Subscription Business Model—Predictive Analytics on Customer Churn. BCP Business and Management, 44(1), 870-876. doi:http://dx.doi.org/10.54691/bcpbm.v44i.4971