Customer Churn Prediction for Subscription-based Business

Choo Peng Tan1, 2, a), Ping Liang Lim2, b) and Choo Kim Tan1, 2, c)

1Centre for Intelligent Cloud Computing, CoE for Advanced Cloud, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia

2Faculty of Information Science and Technology, Multimedia University, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia

a)Corresponding author: cptan@mmu.edu.my

b) pl.lim0624@gmail.com  
c)cktan@mmu.edu.my

**Abstract.** Customer churn prediction has become increasingly important for subscription-based businesses seeking to sustain their growth and maintain an edge over competition. In the current landscape, customer loss has a direct negative impact on any business, including the loss of recurring revenue and the loss of customer lifetime value. In this study, we analyzed basic machine learning methods to predict customer churn using a real-world telecommunications dataset. We developed a detailed data preprocessing pipeline which addressed issues such as missing data, feature engineering, and normalization. After conducting exploratory data analysis (EDA) on the dataset, we trained and evaluated four predictive models (Logistic Regression, Decision Tree, Random Forest, and XGBoost). The model with the best performance was XGBoost, which achieved an accuracy of 85.4% and a ROC-AUC of 89.7%. We found that critical reasons/churn drivers for customers to churn included: 1. Short customer tenure, 2. Higher monthly charges, and 3. Month-to-month contracts. Additionally, we proposed possible recommendations on how churn prediction can be incorporated into existing business intelligence systems to change/renew the customer retention approach, improve personalization, and better manage operational resources. We also addressed challenges, limitations, and future opportunities on predictive analytics, as well as additional opportunities (for example, real-time monitoring or customized analysis on more finely segmented customers), to provide a holistic perspective on this growing field.

# INTRODUCTION

Customer churn is a big concern for subscription businesses because it affects businesses' income stability, forecasting, and planning for the future. Businesses that rely on recurring revenue--such as telecoms, streaming services, and SaaS platforms--are particularly exposed to churn when a customer leaves. Many studies have found that acquiring new customers is significantly more expensive than retaining customers, with some research indicating that the cost of acquiring a new customer could be five times as much. For this reason, reducing churn is a top concern for sustainable business growth.

Predicting customer churn involves measuring customer data to identify behaviors or patterns that may indicate dissatisfaction or indicate high likelihood of service cancellation. Most traditional methods of churn analysis focus on reactive techniques (like customer surveys or other data collected after customers leave), which have limited time-based responsiveness and could not uncover nuanced behavior changes from the customer. So, they are not reliable ways to predict churn that allow businesses to take corrective actions in time. Machine learning (ML) techniques represent a proactive approach to help businesses quickly find at-risk customers even if they don't leave a bad review or send negative feedback, based on historical (and behavioral) data. Machine learning systems can train predictive models based on existing customer characteristics such as demographic information, usage patterns, valuations, and service usage. With this, ML systems can estimate the likelihood of churn for each customer and allow the business to offer targeted intervention before churn happens. If a customer is reported dangerous, firms can offer discounts, packages, or loyalty rewards.

For many firms where machine learning makes sense, the issue is more about implementing the models in their own operations in real-time, and this can be very difficult for certain reasons such as data quality, balance of the classes, and that complex models are often not interpretable. Also, so many of current models are also not aligned to the needs of the businesses have and makes it impossible for decision makers to adequately act on results. Once again, dashboards or other visual approaches are rarely available to provide seamless interactivity to the insights.

The research project will provide a churn prediction system designed for firms using a subscription model. The goal of the project is a churn system providing a combination of classical and ensemble machine learning models (Logistic regression, Decision tree, Random Forest, and XGBoost) using a real-world telecom data set. The study will involve significant data wrangling and exploratory data analysis (EDA), then building, tuning, and assessing machine learning models, while also using traditional metrics (accuracy, precision, recall, F1-score, ROC-AUC) to evaluate predictive performance. The winning model will be the basis for an interactive churn prediction dashboard in the future.

Although accurately predicting customer churn is essential, this study will largely focus on the interpretability of the model and its applicability to businesses. we will be conducting a feature importance analysis, which will indicate the key driving factors of churn, allowing us to develop actionable business strategies. Ideally, we would create a comprehensive and interpretable solution that not only predicts churn but also helps organizations proactively keep their customers. Furthermore, understanding churn is imperative because in an increasingly competitive digital economy, where it is easy to switch digital providers, understanding churn is required.

Organizations that provide telecommunications, Software as a Service (SaaS), or Over-the-top (OTT) platforms like Netflix, Spotify, or Amazon Prime, are under tremendous pressure to provide consistent customer engagement. Predictive churn analysis enables those companies to transition from reactive customer care to proactive relationship management. Any organization can do a better job of not only preventing churn but also tailoring their marketing campaigns, service upgrades, and onboarding processes using data-based indicators. As a result, organizations no longer rely on simple predictive churn analysis, but instead optimize their customer experience, an essential differentiator in the subscription economy.

# Literature Review

In recent years, explainable artificial intelligence (XAI) has gained traction in churn prediction research. Techniques such as SHAP (SHapley Additive exPlanations) have been widely adopted to provide interpretable model outputs and reveal key churn-driving features at both global and individual levels [1]. Studies also highlight the integration of churn predictions into real-time dashboards and CRM systems, emphasizing the importance of usability in business environments [2]. Moreover, hybrid architectures combining XGBoost with neural networks or deep embeddings have shown promise in enhancing prediction accuracy while maintaining interpretability [3]. These advancements suggest that the future of churn prediction lies not only in precision, but in the seamless alignment between model performance and decision-making frameworks.

Customer churn prediction has been widely researched in both the academic and commercial spaces. Researchers have utilized a variety of statistical and machine learning methodologies and techniques in the pursuit of finding customers who will likely churn. Verbraken et al. [4] indicated that predictive churn modeling offers a more precise and proactive outlook than descriptive modeling, delivering cost savings and better targeting for retention campaigns. Logistic Regression was commonly used in the early churn prediction models due to the ease of implementing and interpreting the results as it quantified the likelihood/risk of churn based on independent variables such as customer tenure or billing method. However, Logistic Regression assumes that all relationships between the independent features and dependent outcome variable are linear and cannot accommodate more complex relationships when dealing with larger datasets.

An alternative approach to Logistic Regression type of models was a non-parametric method using a Decision Tree as an improvement upon the previous models. Decision Trees, however, do not assume linear relationships and can model non-linear relationships well but suffer from overfitting again particularly when trained on datasets with noise. A Random Forest model was created to help overcome the overfitting problem, by creating multiple Decision Trees and averaging the output from then to enhance generalization. It addresses imbalanced data more effectively and delivers excellent results with little or no preprocessing.

In recent years, Gradient Boosting algorithms like XGBoost have emerged as the state of the art for churn problem tasks. XGBoost boosts the accuracy of weak learners iteratively while simultaneously controlling overfitting through regularization. Multiple studies demonstrate benignly that XGBoost outperforms Random Forest in the majority of classification tasks, particularly with the appropriate tuning through hyperparameter optimization implication such as Grid Search.

Several studies have also suggested hybrid and ensemble models which combine several base learners to further improve performance. For example, Chen et al. [5] proposed a model stacking with Random Forest and Logistic Regression, which resulted in improved predicted accuracy compared to either model alone. In addition, Tang et al. [6] also explore using deep learning combined with boosting models to extract better features within larger scale datasets.

Despite these advances, there remain challenges - real-world churn datasets are usually very imbalanced (with sometimes only a small percentage of customers actually leaving); thus leading to unwanted implications making model training and increasing false negative rates. Moreover, model interpretability is essential to business usage. For example, despite more accurate predictions, black-box models such as XGboost are often challenging to account for to clientele. Thus, the increasing attention surrounding explainable machine learning and tools such as SHAP and LIME seeks to bridge a gap between accuracy and transparency.

In addition to model performance, this year, literature and researchers have focused on the importance of embedding churn predictions into the business process with the use of visual analytics tools and customer segmentation schemas. For instance, analytical dashboards with business intelligence tools (e.g., Power BI, Tableau) allow the retention teams to quickly filter high-risk customers by tenure or contract type in order to carry out targeted actions.

Recently, Adiputra and Wanchai [7] developed a new weighted ensemble model that combined statistical and tree-based methodologies to address data set imbalance and model robustness. They continued to strongly endorse the notion that no one model can be relied upon in all situations, and that stacking and blending can produce more stable results. Other researchers have considered the use of survival analysis where prediction of time-to-churn, not only provides a binary outcome, but also indicates likely time frame when a customer may churn. While many of the models listed above emphasize short term churn prediction, an increasing number of studies are exploring a long-term loyalty modeling approach in which not only cancellation events are observed, but also include rates of downgrading, late payers, and drop-off usage. In this way, long-term loyalty modeling gives more insight into customer decision making over time.

Additionally, regulatory concerns related to automated decision-making have intensified efforts to build models that are not only technically accurate but also explainable and fair. For example, transparency techniques such as with SHAP (SHapley Additive exPlanations) allow analysts to quantify how much each feature contributes to a prediction, reducing doubt and increasing trust for compliance with data regulations.

In addition to algorithmic improvements, researchers have also examined feature selection methods in order to improve churn prediction. There are three types of methods (i.e., wrappers, filters, and embedded) that have been utilized to identify the most important attributes, reducing (a) overfitting, and (b) reducing the computation time. Furthermore, interest in time-series analysis and customer journey modeling continues to thrive, with churn risk calculated dynamically as the customer's behavior evolves over time. This is a change from static models to dynamic systems that reflect real-world user behavior. In some cases, researchers take this further and include psychological and behavioral theories in their model construction.

Organizations have started recognizing that churn cannot be merely thought of as a transactional process, because adding psychological and behavioral frameworks and theories to machine learning can help businesses define motivational triggers that drive customer decisions. For example, the Technology Acceptance Model (TAM) and Expectation-Confirmation Theory (ECT) can each add depth to the model, while still intertwining machine learning and future churn predictions.

# Methodology

The research started by acquiring data from an open-source telecommunications dataset of customer profiles that included usage patterns, contract types, payment mechanisms and churn. It consisted of 7,043 customer entries, with 21 attributes, and the 'Churn' variable serving as the binary target label. The preprocessing stage involved the elimination of missing values, normalizing numerical variables, and converting categorical variables. The TotalCharges column, in particular, had some missing values which were imputed using its median. Gender, SeniorCitizen and other nominal variables were transformed using label encoding and multi-class categorical features including InternetService, Contract, and PaymentMethod were transformed using duple-one-hot encoding.

Due to various machine-learning models and algorithms will use the categorical variable as input, it is important to convert the variables without introducing bias. Normalization was used with numerical attributes, such as MonthlyCharges and Tenure using MinMaxScaler, to keep scale uniformity across features. The treatment of outlier values was performed using the Interquartile Range (IQR) approach compared to the features' standard deviation, as they contained values that, in some cases, were many multiples of averages of values with similar inputs from other records. Feature engineering was used to separate tenure groups (for example 0–12 months, 13–24 months, etc.) and composite indicators such as "TotalPaid" being calculated from the product of MonthlyCharges multiplied by Tenure.

Exploratory Data Analysis (EDA) was done using visualization libraries like Seaborn and Matplotlib. The exploratory data analysis techniques used this include histograms, bar plots, and correlation matrices to help identify things like the churn risk associated with monthly contracts versus long-term engagements and the additional risk posed by customers paying by electronic check. After EDA, modeling was done in four stages by training four classifiers: Logistic Regression, Decision Tree, Random Forest, and XGBoost. The data set was divided into 80% for training and 20% for testing. A Grid Search with a 5-fold cross-validation was done to tune hyperparameters and find the optimal value of the learning rate, depth of tree, and number of estimators.

As further data quality check, additional checks were done through descriptive statistics and visual inspections. Skewness and kurtosis values were analyzed to check for distributions that did not conform to expected distributions, and when necessary, were corrected using a log-transformation on the variable of concern. For example, the variable TotalCharges was positively skewed, and after checking, a log-normal transformation was done to permit use of algorithms that assume their data are normally distributed or at least consider the assumptions when estimating their coefficients.

As part of training the final models, a baseline dummy classifier (i.e., one that predicted the majority class, non-churn) was included to determine the actual incremental benefit afforded by machine learning models relative to a dummy baseline. The dummy model was also used to produce benchmarks for the improvements made in recall and F1-score. All four main models showed significant improvement with respect to the dummy classifier, providing adequate justification for the complexity and training time of the final models. The performance of models was examined through a variety of classification metrics. Accuracy was calculated, but it was supplemented by recall and F1-score to take into consideration the imbalanced nature of the dataset. ROC-AUC scores were also calculated so that benchmark assessment of classifiers’ abilities to differentiate churners from non-churners could be assessed across various thresholds.

Experiments were run in a Python environment using Jupyter Notebook, using various libraries including Pandas, NumPy, Scikit-learn, XGBoost, and Seaborn. Modularization of the code and consistent assessments allowed for reproducibility and fair comparisons of results. The workflow also had one step that involved balancing the input data, since the original classes were imbalanced, with only about 26% of the data being churned customers.

As a way to help balance the data, we tested Synthetic Minority Over-sampling Technique to synthetically create samples of the minority class. There were model assessments which showed that some encoding methods, like the ensemble tree-based model XGBoost, handled classification balance better, which made resampling relatively unnecessary. Additionally, the feature selection process also eliminated unnecessary features to reduce noise and collinearity through mutual information and correlation. For example, TotalCharges and MonthlyCharges correlated highly with tenure and having all three on there without adjustment would lead to multicollinearity. Not only did we prune those two features with mutual information but we relied on feature importance scores from tree-based models to prune additional low-value features to help reduce model complexity and improve model generalization. We also assessed a second ensemble method by combining Logistic Regression with Random Forest in a soft-voting classifier. This method looked promising in the initial tests but ultimately was outperformed by XGBoost in terms of recall and the ROC-AUC metric.

The flowchart summarizes the sequential process adopted in this study, from data collection through preprocessing, exploratory data analysis (EDA), model development, and final evaluation (refer to Figure 1). This structured approach ensures clarity and reproducibility in churn prediction modeling.

Data Collection

Preprocessing

EDA

Evaluation

Recommendation

Model Training

**FIGURE 1.** Flowchart of methodology pipeline

# Results and Discussion

The four models showed a range of results across the accuracy, recall, and classification results. While Logistic Regression is very interpretable, it has a moderate performance as it correctly classified 79.3% of instances and achieved a ROC-AUC of 81.2%. It struggled to generalise in more complex circumstances because it treated all relationships as linear and did not consider any feature interactions. Decision Tree lost a little performance for accuracy at 78.5% and ROC-AUC at 79% because of overfitting. However, the decision paths were informative, providing the full interpretability benefits for identifying the rule based segments for churn.

Random Forest slightly improved performance again where we recorded and accuracy of 83.1% and ROCs of 86.5%. Random Forest is less likely to overfit and captures more than one decision boundary for customer segments. In our feature importance plots, we noticed redundancies between correlated attributes such as TotalCharges and MonthlyCharges. The unsupervised algorithms did not consider certain aspects of customer churn in the feature correlations.

As reported in the beginning of this pragmatic report, XGBoost was the best performer with an accuracy of 85.4% and the best ROC-AUC at 89.7%. XGBoost provided very high predictive power and at the same time possible generalizability. The feature importance analysis of XGBoost identified Tenure, contract type, monthly charges, and online security as the most important features (refer to Figure 2). The feature importance plot from the XGBoost model shows that Tenure, Contract Type, and Monthly Charges are the most influential features in predicting churn. These insights align with business logic and help focus retention strategies on short-tenure customers or those with monthly contracts.

A graph of a number of orange bars

AI-generated content may be incorrect.

**FIGURE 2.** Feature importance plot for XGBoost

From a business perspective, these features certainly made sense, which provided credibility to the insights. In addition, to assess model robustness, the results were conducted with different train-test splits (70-30, 90-10), and XGBoost consistently outperformed the other models in all partitions, which highlights consistency in model performance. It could be assumed that the model is capable of generalizing to unseen data, given this level of stability across different data splits. The precision-recall curve of XGBoost also demonstrated a good balance, with just a slight compromise to recall when optimizing for precision, which is critical in this case because we want to minimize the likelihood of false negatives when detecting churn. A further insight was uncovered around customer segments.

When plotting churn risk by tenure band, we found that customers between 0-6 months were the least predictable, indicating that the newest customers may have more to do with the onboarding experience than we understood and could lead to churn in future. Conversely, customers with tenures of 6-12 months exhibited more stable behaviour, which indicates that it could be a transitional point for some customers where the transition either strengthens or erodes loyalty. This could inform related communication plans at onboarding and activate early loyalty behaviour.

The ROC curve comparison indicates that XGBoost achieved the highest AUC value among all models, followed by Random Forest (refer to Figure 3). This demonstrates XGBoost's superior ability to distinguish between churn and non-churn classes, making it the most reliable model in this study. The ROC curve for each of the models also demonstrated the trade-offs between sensitivity and specificity in context. Although Logistic Regression demonstrated stable sensitivity, its area under the curve tapered off earlier than the Random Forest and XGBoost suite, indicating that it has less discrimination power and resulting performance drop at higher thresholds.

# A diagram of a graph AI-generated content may be incorrect.

**FIGURE 3.** ROC curve comparison for all models

The confusion matrix (Figure 4) shows that XGBoost performs well in identifying both churners and non-churners, with a relatively low number of false negatives. This suggests the model is effective at minimizing missed churn cases, which is critical in proactive retention strategies.

A yellow and purple squares

AI-generated content may be incorrect.

**FIGURE 4.** Confusion matrix of XGBoost

The model comparison, depicted in Figure 5, indicates that XGBoost was the best performer on all metrics. Its capacity to capture complex non-linear patterns combined with its robustness towards noisy data contributed most of its power. Training time-wise, Logistic Regression and Decision Trees had the lowest training time (<1 second). Random Forest and XGBoost required relatively longer training times. For example, XGBoost, including cross-validation, took approximately 8-10 seconds for full training. While this is a trade-off compared to the performance of XGBoost in accuracy and stability, it is more acceptable in environments where retraining a model occurs periodically rather than in real-time.

A graph showing different colored squares

AI-generated content may be incorrect.

**FIGURE 5.** Model accuracy comparison

A second beneficial characteristic of tree-based models is the ability to expose non-linear interactions. For example, the churn probability escalated nearly exponentially when both monthly charges and contract type (length) were high (short term) - an aspect that is unlikely to be captured in any linear model. This highlights our use of ensemble learning for high-dimensional, non-linear problem spaces.

Moreover, the analysis also provided opportunities to think about predictive churn modeling as part of a larger customer lifecycle management plan. By combining predictive modeling insights with automated marketing triggers or CRM alert systems, companies could develop customer reassurance and retention campaigns just at the point that risk to leave a contract appears, radically shifting organizational response from a reactive stance to an anticipatory one.

From an overall business strategy standpoint, organizations should be focusing on the customers with short tenure and monthly contracts, so that individualized, personalized marketing efforts and assurances, or relics to stay, could take place to mitigate their risk of leaving - for example, organizations should look to send long-term incentives to customers at an early stage of tenure. Similarly, intentionally bundling service, such as technical support or online security services, creates additional value for customers, and then may reduce churn as perceived value rises.

In summary, predictive churn modeling could generate a churn probability score, enabling modeling teams to visually show a churn probability score in a business dashboard, and then enabling proactive retention activity with the department assigned to the periodic retention of data quality for customers that were ranked as a high risk in almost real time. With the classification results, the confusion matrices indicated the degree that the XGBoost generated classification model had a superior preference to minimize the rate of false negative churn, when compared to the other predictive models for churn predictions. Therefore, there is a benefit to using XGBoost because it gave firms an improved predictive model to better retain at risk customers. This is especially relevant in a startup churn prediction and retention service. In a competitive context where failure to identify a at risk customer should be noted and signaled by account owners, such a failure could have dramatic financial implications to the organization due to leaving the customer for another service company.

# CONCLUSION

Overall, this research shows that machine learning and, more specifically, ensemble techniques, such as XGBoost, can be useful for predicting customer churn for subscription-based businesses. Leveraging structured data and applying preprocessing, feature engineering, and model tuning techniques will allow us to make highly accurate predictions. Of the four models that we tested, XGBoost had the best performance, providing a balance of predictive accuracy and interpretability for practical use. The key predictors of churn—customer tenure, contract type, billing method, and whether support services were involved—followed business logic and provided intrinsically useful information to be used to trigger intervention.

However, while promising, the research is limited. We used a single dataset from the telecommunications domain, which may limit the possible generalization of this study’s findings to different domains. The research also only focused on structured data, with rich data that could have come from unstructured data, including customer feedback or customer chat logs being ignored.

Future work may include combining or aligning the text analytics and sentiment analysis with structured data; essentially creating a multidimensional system for predicting customer churn. Furthermore, if businesses were to develop a real-time dashboard in conjunction with their CRM system, the opportunity exists to more actively, and accurately monitor and react to churn risk. If the predictive modeling could be applied across multiple channels, it would almost certainly enhance the predictive power currently found in many businesses, particularly if it extends to social media, customer service logs, and so on.

This research provides a contribution to the field of customer analytics both in a technical sense and in a strategic sense that is applicable for businesses. The research provides a strategic approach framework that businesses could use to gain a better understanding of how they can better churn, retention, and even customer lifetime value through intelligent, data-driven decision making. There is a plethora of practical applications to this paper that goes beyond churn classifications. The modelling framework in this paper could be adapted to suit other classification challenges extraordinarily important to any business such as predicting customer service upgrades, predicting cross-sell probability, or even customer satisfaction.

Moreover, the suggested approach allows for modular integration with CRM and marketing automation programs. Most importantly, if built into the system architecture, a retraining process should be included to monitor and maintain adaptability and performance. A new model can then be developed by retaining the original and using the new data. The business will be able to manage evolving customer behaviours, and account for any degradation in model predictions over time.

Lastly, the organizations who deploy such systems should evaluate the ethical implications of automated decision-making. This means thinking about transparency, customer consent, and fairness in model outcomes, especially in contexts where decision-making may result in customer exclusion from services or benefits. Additionally, integrating explainable AI in their systems will allow the customer service teams to rationalize their interventions, while fostering user trust. Machine learning for churn prediction has tactical and strategic benefits. It allows marketing, operations, and data science teams to operate on common retention goals. Future research could also consider how to deploy these models as APIs on cloud services or integrate them into mobile applications for customer self-service and retention diagnostics. Ultimately, it will be the combination of accurate prediction, interpretability, and business integration that will provide the foundation for successful churn analytics in real world deployments.

# References

1. H. Kim and D. Lee, “Deep hybrid models for churn prediction in subscription services,” Expert Systems with Applications 216, 119416 (2023).
2. J. Zhang, B. Liu, and Y. Zhang, “Real-time customer churn prediction with interactive dashboards: Bridging model performance and user needs,” Information Systems Frontiers 24, 67–81 (2022).
3. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Advances in Neural Information Processing Systems 30, (2017).
4. T. Verbraken, W. Verbeke, and B. Baesens, “Profit Optimizing Customer Churn Prediction with Bayesian Network Classifiers,” *Intelligent Data Analysis*, vol. 18, pp. 3–24, (2014).
5. F. Chen, Y. Zhang, L. Wang, and H. Li, “Customer churn prediction based on stacking model,” in 2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL), (IEEE, 2023).
6. Q. Tang, G. Xia, X. Zhang, and F. Long, “A customer churn prediction model based on XGBoost and MLP,” in 2020 International Conference on Computer Engineering and Application (ICCEA), (IEEE, 2020).
7. N. M. Adiputra and P. Wanchai, “Customer churn prediction using weight average ensemble machine learning model,” in 2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE), (IEEE, 2023).