Predict, Prevent, Retain: Machine Learning for Customer Retention

Maureen Ekwebelem
Fordham University
New York, NY, USA
ORCID: 0009-0006-7729-0829

***Abstract*—Customer churn is a critical challenge in the telecom industry, where retaining existing customers is far more cost-effective than acquiring new ones. This project explores how machine learning can be used to predict and prevent customer churn by analyzing behavioral patterns from a dataset of 3,150 telecom users. After data cleaning and exploratory analysis, this project identified key drivers of churn- customers who churned had on average higher frequency of complaints, higher call failures, and lower usage of the telecom services. To address class imbalance, SMOTE was applied, allowing the models to better detect churners. Several classification models were tested, with Random Forest and XGBoost delivering the strongest performance in terms of recall, precision, and F1-score. Beyond predictive accuracy, the project emphasizes the importance of using model insights to guide flexible, human-centered retention strategies. When applied systematically, machine learning can help telecom companies not just predict churn but take meaningful steps to reduce it.**

Keywords—Machine learning, classification models, customer retention, predictive modeling

I. INTRODUCTION

Customer churn, the rate at which users leave a service, is a major challenge in the telecom industry. It leads not only to revenue loss but also drives up customer acquisition costs, making it a costly issue for companies to ignore [1]. Retaining existing customers is significantly more cost-effective than acquiring new ones, which is why the ability to identify at-risk customers early has become a key competitive advantage [2]. Even small improvements in retention can translate into meaningful long-term business value.

This project focused on a dataset of 3,150 customers from an Iranian telecom company, available through the UC Irvine Machine Learning Repository [3]. The dataset contains user activity over the first nine months and labeled churn status at the twelve-month mark. This time-based structure supported a supervised learning approach, allowing models to predict churn before it occurs and uncover behavioral patterns that may signal future customer loss.

As companies gather increasingly large volumes of customer data, machine learning has become a porful tool for surfacing patterns and insights that traditional methods often miss [4]. Rather than reacting after a customer has already left, these models enable businesses to take a proactive stance, identifying signs of churn earlier and gaining a deeper understanding of why customers disengage. When implemented effectively, predictive modeling not only helps reduce churn but also contributes to building stronger, more meaningful relationships with customers.

In this analysis, I applied supervised classification models to predict customer churn in the telecom sector, framing the problem as a binary classification task [1]. The goal is to evaluate how well classification models such as Logistic Regression, Random Forest, Naïve Bayes, and XGBoost can predict customers who will churn, and to understand which behavioral features most influence churn outcomes.

The insights gained from this project contribute to a broader understanding of customer behavior and offer practical value for developing data-driven strategies aimed at improving retention and customer satisfaction.

II. APPROACH

This project followed a clear and structured process, beginning with data preprocessing to prepare the dataset for analysis. This was followed by exploratory data analysis (EDA) to uncover patterns, trends, and potential relationships between features and churn. Given the imbalance in class distribution , where churned customers represented a minority, techniques were applied to address this issue and ensure fair model training. Finally, a range of predictive classification models were built and evaluated to assess their effectiveness in identifying churn risk.

A. Data Preprocessing

The first step was preparing the raw dataset for modeling. Fortunately, the dataset had no missing values, which saved time and reduced the need for imputation. Still, a few adjustments were necessary to get the data in shape:

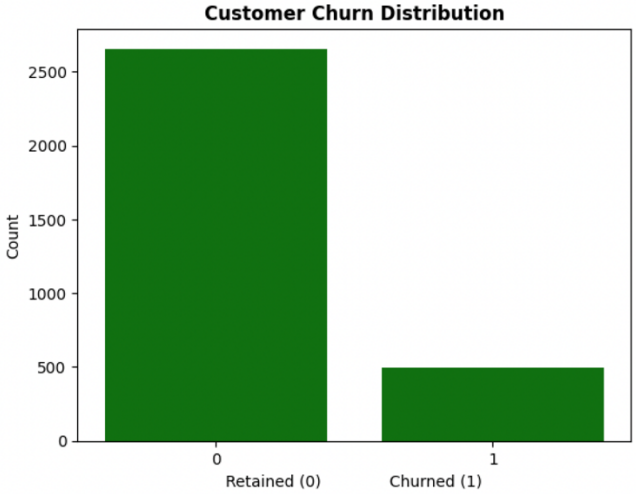
- **Monthly averaging:** Since many of the usage-related features (e.g., total charge, total minutes used) were cumulative over 9 months, I converted them into monthly averages to normalize across the time window. This improved interpretability and allowed for more meaningful comparisons across models.

- **Feature scaling:** I standardized the continuous variables so that models sensitive to feature magnitude, for example Logistic Regression, would perform reliably and treat all features on an equal playing field.

- **Categorical Encoding:** All categorical features were converted to numerical format using label encoding to ensure compatibility with scikit-learn models.

B. Exploratory Data Analysis

I examined the distribution of the target variable (churn) to understand potential imbalances in the data. As shown in the bar chart (Fig. 1), only 15.7% of customers churned, while 84.3% were retained, revealing a significant class imbalance.




```
Churn Rate:
 Churn
0    0.842857
1    0.157143
Name: proportion, dtype: float64
```

Fig. 1. Customer churn distribution: majority retained (84.3%), minority churned (15.7%).

This imbalance creates a challenge for predictive modeling, as many classification algorithms tend to favor the majority class by default. To address this, I planned to apply stratified test-split and SMOTE (Synthetic Minority Oversampling Technique) for training which I will discuss further on.

Using the Seaborn a correlation heatmap (Fig. 2) to explore how different features relate to one another and to the churn variable. Examining the heatmap gave us an early sense of which features may be influencing customer behavior and provided a foundation to guide the rest of our analysis.

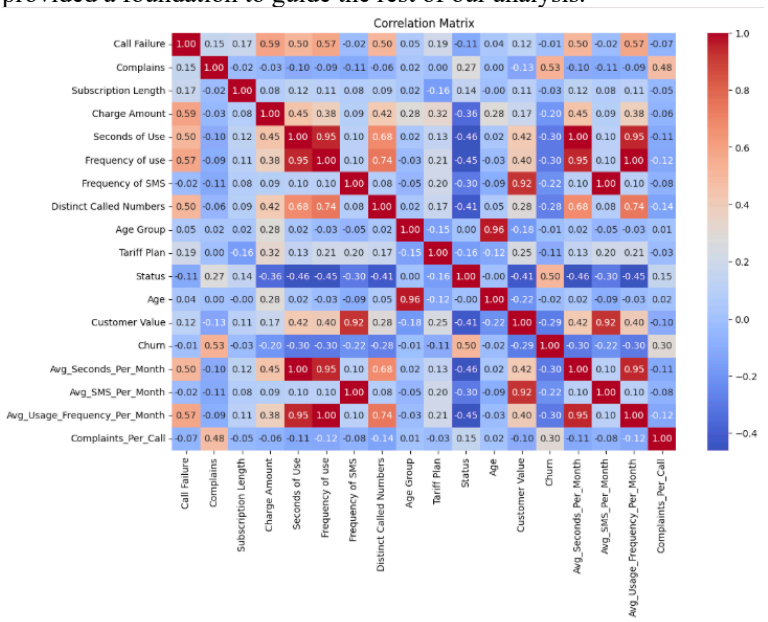


Fig. 2. Correlation matrix generated with Seaborn to identify relationships among features.

The heatmap made it easier to visualize which features were closely related to churn and which aren't contributing as much (Fig. 2). Based on this, prior to modeling, I removed features like Age, Frequency of SMS, and Distinct Called Numbers. While they might seem useful on the surface, they didn't show strong correlations or add value to the model's performance. This step helped simplify the dataset proceed in the analysis with more strongly correlated features for predicting churn.

Churn	Complains	Customer Value	Seconds of Use	Frequency of use
0	0.015443	0.124849	0.129076	0.130977
1	0.404040	-0.669644	-0.692318	-0.702515

Churn	Subscription Length
0	0.014071
1	-0.075472

Fig. 3. Comparison of mean feature values for churned vs. retained customers.

In order to understand customer behavior, I compared feature averages between churned and retained customers (Fig. 3). Churned customers on average reported significantly more complaints (0.404 vs. 0.015), suggesting that dissatisfaction plays a major role in churn. They also on average had much lower customer value to the business (-0.670 vs. 0.125), shorter subscription lengths (-0.075 vs. 0.014), and lower usage across both frequency and time duration (-0.702 and -0.692 vs. 0.131 and 0.129, respectively).

After reviewing the feature comparisons (Fig. 3), I decided to focus specifically on Customer Value and Seconds of Use, as they showed strong differences between churned and

retained customers. To better visualize these differences, I used boxplots to examine the distribution of each feature across the two groups (Fig. 4).

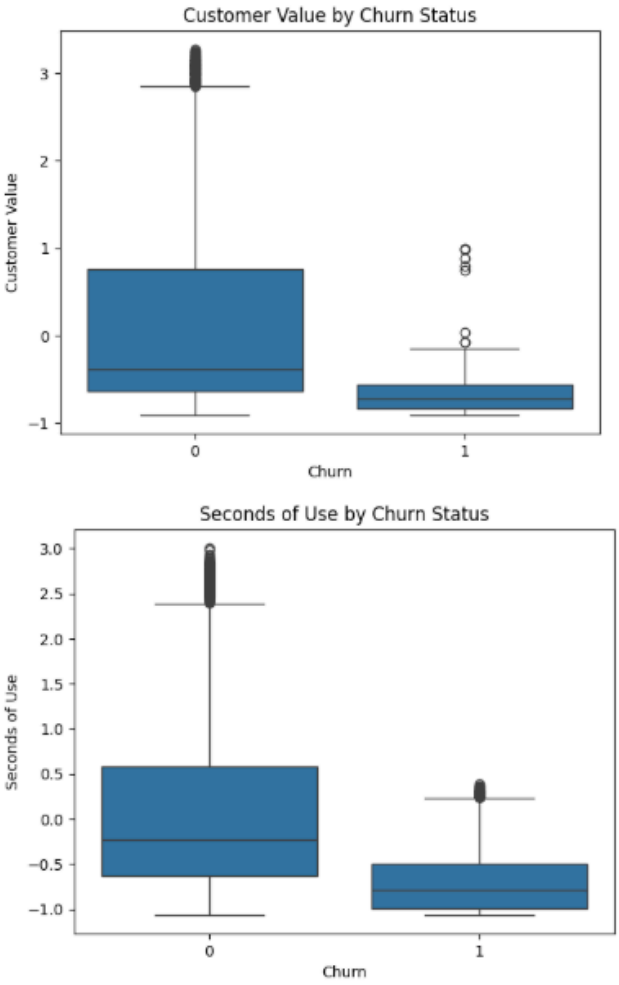


Fig 4. (top) Customer Value by Churn Status; (bottom) Seconds of Use by Churn Status. Values are standardized using z-scores, where 0 represents the overall customer average.

Although customers who churned tended to have lower value to the company, as shown in Fig. 4, this does not necessarily mean they were beyond retention. Instead, the insight reinforces the need to identify at-risk users early and provide opportunities for re-engagement before churn becomes final.

In conclusion of the exploratory data anlalysis, notable findings included:

- Churned customers had more complaints (0.404 vs. 0.015), shorter subscription lengths, lower customer value, and used the service less overall-clear indicators of dissatisfaction and/or disengagement.
- Retained customers showed the opposite trend, with higher average usage, longer subscription durations, greater value to the company, and overall stronger engagement.

These early patterns pointed to a connection between lower engagement and churn, though it's worth noting that low usage alone doesn't always reflect dissatisfaction, it could also be due to lifestyle or external factors. Still, when combined with other signals, disengagement appeared to be a useful indicator to consider during predictive modeling.

III. Modeling & Comparison

A. *Handling Class Imbalance*

One of the most critical considerations in this project was class imbalance, as only about 15% of customers in the dataset had churned. Recognizing and addressing class imbalance early in the process was essential to building fair and effective models. To address it, I applied SMOTE (Synthetic Minority Over-sampling Technique) only to the training set. SMOTE helps balance the data by creating new, synthetic examples of the minority class based on the ones that already exist, making it easier for the model to learn a fairer decision boundary [5]. This played a critical role in improving recall and F1-score, helping the model better identify churners without sacrificing overall performance. Stratified train-test splitting helped to maintain the churn vs non churned ratio in both sets to ensure consistency and fair evaluation.

After preprocessing and EDA, I tested four classification models:

- Logistic Regression
- Random Forest
- Naive Bayes
- XGBoost

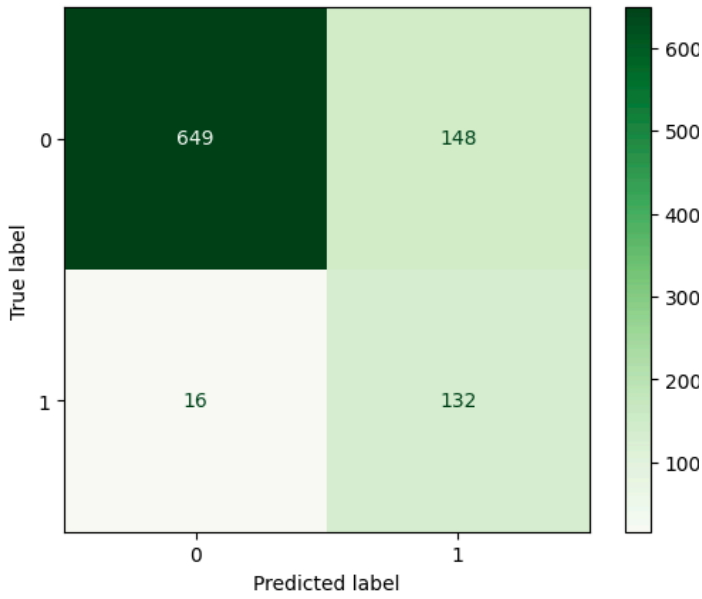
For evaluation, I focused on accuracy, precision, recall, and F1-score, since the main goal was to identify real churners without generating too many false alarms.

B. *Model Evaluation & Interpretation*

Each model had its own strengths, with some performing better overall than others depending on how they balanced precision, recall, and accuracy.

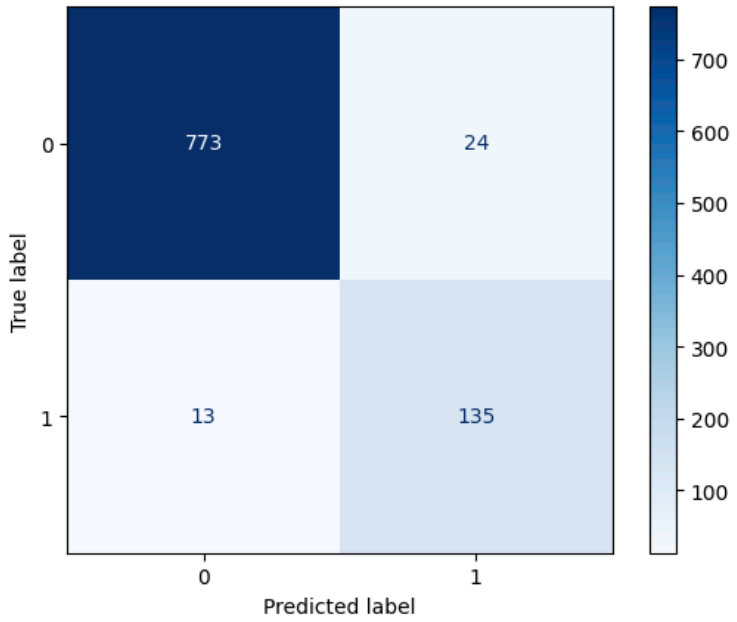
- **Logistic Regression:** Logistic Regression achieved an accuracy of 83%, with a recall of 0.89 and precision of 0.47, leading to an F1-score of 0.62 (Fig 5). This model was effective at identifying most customers who churned, as shown by the high recall score. However, the trade-off was a relatively low precision, meaning that it also predicted churn in many customers who actually stayed. In other words, while the model was good at catching

churners, it tended to raise too many false alarms-flagging loyal customers as churn risks. This could potentially lead to unnecessary retention efforts or customer outreach in real-world use.



	precision	recall	f1-score	support
0	0.98	0.81	0.89	797
1	0.47	0.89	0.62	148
accuracy			0.83	945
macro avg	0.72	0.85	0.75	945
weighted avg	0.90	0.83	0.85	945

Fig. 5. Logistic Regression model results. (top) Confusion matrix showing prediction breakdown. (bottom) Classification report highlighting high recall (0.91), low precision (0.47), and overall accuracy of 83%.

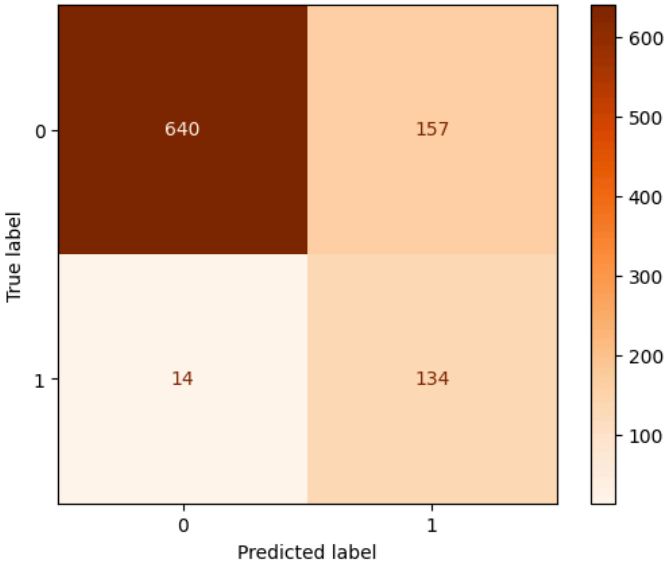


	precision	recall	f1-score	support
0	0.98	0.97	0.98	797
1	0.85	0.91	0.88	148
accuracy			0.96	945
macro avg	0.92	0.94	0.93	945
weighted avg	0.96	0.96	0.96	945

Fig. 6. Random Forest model results. (top) Confusion matrix showing prediction breakdown. (bottom) Classification report highlighting high recall (0.91), strong precision (0.85), and overall accuracy of 96%.

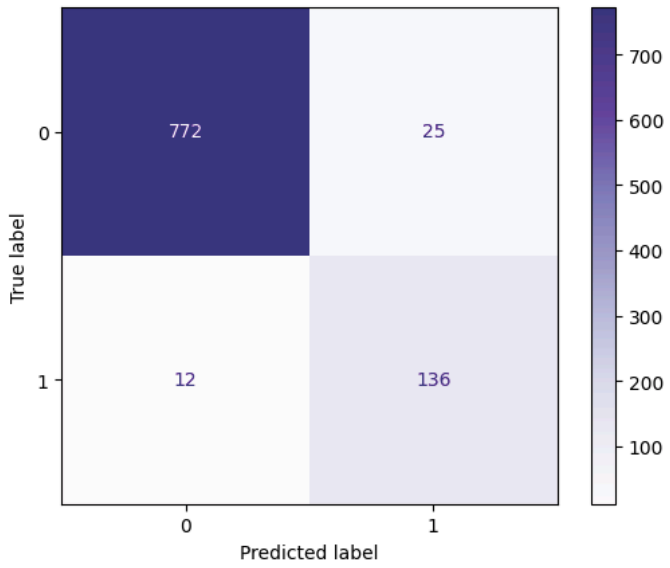
- **Random Forest:** Random Forest showed excellent performance, with a high accuracy of 96%, a precision of 0.85, recall of 0.91, and a strong F1-score of 0.88 (Fig. 6). This model not only correctly identified most churners but also did so with fewer false positives compared to Logistic Regression or Naive Bayes. The balance between precision and recall suggests that Random Forest was both aggressive and accurate, successfully predicting churn without overly misclassifying non-churners. Its strong performance across all metrics makes it a reliable choice for churn prediction tasks.

- **Naïve Bayes:** Naive Bayes had an accuracy of 82%, recall of 0.91, precision of 0.46, and an F1-score of 0.61(Fig. 7). While its accuracy was slightly lower than Logistic Regression, it had an equally high recall, meaning it was also very good at identifying customers who were likely to churn. However, similar to Logistic Regression, the low precision indicates that many of these churn predictions were incorrect. This could be problematic in a business setting, as it might result in misdirected retention resources or unnecessary intervention strategies.



	precision	recall	f1-score	support
0	0.98	0.80	0.88	797
1	0.46	0.91	0.61	148
accuracy			0.82	945
macro avg	0.72	0.85	0.75	945
weighted avg	0.90	0.82	0.84	945

Fig. 7. Naïve Bayes model results. (top) Confusion matrix showing prediction breakdown. (bottom) Classification report highlighting high recall (0.91), low precision (0.46), and overall accuracy of 82%.



	precision	recall	f1-score	support
0	0.98	0.97	0.98	797
1	0.84	0.92	0.88	148
accuracy			0.96	945
macro avg	0.91	0.94	0.93	945
weighted avg	0.96	0.96	0.96	945

Fig. 8. XGBoost model results. (top) Confusion matrix showing prediction breakdown. (bottom) Classification report highlighting high recall (0.91), strong precision (0.85), and overall accuracy of 96%.

- *XGBoost:* XGBoost matched Random Forest with an impressive accuracy of 96%, and achieved precision and recall scores of 0.85 and 0.91, respectively, yielding an F1-score of 0.88 (Fig. 8). Among all the models, XGBoost had the highest recall, which means it was the most effective at detecting nearly all churners in the dataset. It also maintained a strong precision, suggesting it made fewer incorrect predictions about loyal customers. This makes XGBoost a powerful and well-rounded model for churn prediction, especially when the goal is to minimize the risk of missing potential churners.

C. Interpretation & Insights

Across all models, the features most strongly associated with churn were:

1) Number of complaints
2) Call failures
3) Short subscription lengths
4) Lower usage (minutes and charges)

These findings aligned closely with both the patterns uncovered during exploratory data analysis and the expected business logic: customers who felt underserved, frustrated, or disengaged were significantly more likely to churn. This was evident not only in the visualizations (such as boxplots and correlation heatmaps) but also in the model results, where dissatisfaction and low engagement consistently predicted churn.

Among the classification models tested, Random Forest and XGBoost consistently delivered the strongest overall performance, striking a solid balance between high recall (capturing most churners) and good precision (limiting false positives). This balance is particularly important in real-world

business settings, where overly aggressive churn predictions can lead to unnecessary or misdirected retention efforts. Notably, these results align with trends observed in similar machine learning projects. For instance, prior churn analyses have also identified XGBoost as a top-performing model, likely due to its ability to manage class imbalance and capture subtle behavioral patterns [1].

These findings also echoed patterns seen during exploratory analysis. Customers who felt underserved or disconnected were more likely to churn- a trend supported by both the statistical results and business logic [6]. Random Forest and XGBoost stood out in their ability to reflect these patterns accurately.

IV. Conclusion

This project highlighted how machine learning and strategic data analysis can support smarter, data-driven decisions around customer retention. By combining technical modeling with business insights, the analysis showed how predictive tools can be used not just to identify risk, but to take meaningful action. To reiterate, here are a few key insights from the analysis:

- **Complaints, low usage, and call failures are strong signals that a customer is at risk of leaving.** These behavioral patterns suggest dissatisfaction and low engagement, which frequently preceded churn in the dataset.
- **Class imbalance** typically needs to be addressed in churn prediction models. Applying **SMOTE**, along with **stratified sampling**, led to more balanced training data and significantly improved recall and F1-score, especially for models like Logistic Regression and Naive Bayes.
- **Random Forest and XGBoost** stood out as the best-performing models. Both delivered high accuracy, strong recall, and good overall precision, making them solid choices for real-world deployment where catching churners early is key.

Most importantly, this project emphasized that models are only as useful as their ability to inform decisions. High-performing metrics are helpful, but the true value comes from being able to explain *why* a customer might churn and then using that insight to take meaningful action. A model that is interpretable and aligned with business logic creates more opportunities for teams to build trust and act early.

For retention strategies informed by these models, the telecom company could consider the following:

- **Highlight underused features:** Send usage-based tips or reminders to customers, especially those with lower engagement. These nudges may encourage customers to get more out of their plan and improve perceived value.
- **Offer conditional perks:** Encourage re-engagement through incentives- like $5–$20 gift cards from partners such as Amazon or Target- but only when a customer takes a specific action (e.g., trying a different plan or increasing their usage).
- **Personalize outreach:** Follow up with at-risk customers in a way that feels genuine and human. A well-timed message that acknowledges their experience can strengthen customer relationships and reduce churn.

In practice, these strategies don't just reduce churn- they build loyalty and improve overall customer satisfaction. While these are just a few potential solutions, the company's insights and strategy team could expand on them to develop more personalized, data-informed approaches. When applied thoughtfully, machine learning can do more than predict churn- it can help companies respond in ways that make customers feel heard, supported, and ultimately, more likely to stay.

While the models performed well overall, it's important to acknowledge a few limitations. Not all low usage reflects dissatisfaction- sometimes, it's just life. That's why retention strategies need to be flexible and human centered. Instead of assuming a customer is unhappy, churn risk predictions can be used as a signal to start a conversation- offering support, options, or simply checking in.

AUTHORS AND AFFILIATIONS

This project was completed by Maureen Ekwebelem as part of the Master's in Data Science program at Fordham University's Graduate School of Arts and Sciences. I led the analysis, modeling, and writing for this paper as an individual submission for CISC 5800 Machine Learning during Spring 2025.

References

[1] Ahmad, A.K., Jafar, A. & Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. *J Big Data* **6**, 28 (2019). https://doi.org/10.1186/s40537-019-0191-6.

[2] Zhengmeng, C., Malik, M., Hussain, M., & Hussain, S. (2024). Exploring customer retention dynamics: A comparative investigation of factors affecting customer retention in the banking sector using mediation-moderation approach. *Heliyon, 10*(19), e36919. https://doi.org/10.1016/j.heliyon.2024.e36919.

[3] Iranian Churn [Dataset]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5JW3Z.

[4] Sarker IH. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. SN Comput Sci. 2021;2(5):377. doi: 10.1007/s42979-021-00765-8. Epub 2021 Jul 12. PMID: 34278328; PMCID: PMC8274472.

[5] Gurcan F, Soylu A. Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. Cancers (Basel). 2024 Oct 8;16(19):3417. doi: 10.3390/cancers16193417. PMID: 39410036; PMCID: PMC11476323.

[6] Xie C, Jin J, Guo X. Impact of the critical factors of customer experience on well-being: Joy and customer satisfaction as mediators. Front Psychol. 2022 Sep 15;13:955130. doi: 10.3389/fpsyg.2022.955130. PMID: 36186283; PMCID: PMC9521495.