Predicting Customer Churn for a Telecom Company

By
Garrett JA Hass

1. Introduction

a. Problem Statement

In the dynamic landscape of the telecommunications industry, customer churn presents a significant challenge to businesses seeking sustained growth and profitability. Our aim is to develop a robust predictive model that effectively identifies customers at risk of churn in an Iranian telecom company. With an emphasis on achieving a balance between the ROC-AUC and F1 scores, this project seeks to provide the company with an actionable solution to proactively address customer churn. The imbalanced nature of the churn data further adds complexity to the task, requiring us to devise strategies that ensure accurate identification of churn instances while minimizing false positives.

By harnessing the power of data science and predictive modeling, we intend to empower the company with insights that enable timely and targeted interventions, ultimately fostering customer retention and business success.

b. Criteria for Success

The success of this project aimed at predicting customer churn for the Iranian telecom company hinges upon the following criteria:

1.High ROC-AUC and F1 Scores: The chosen predictive model must achieve impressive ROC-AUC scores, indicating its proficiency in distinguishing between positive and negative instances. A high F1 score is equally pivotal to ensure a balanced trade-off between precision and recall, effectively capturing the true churn cases while minimizing false positives and false negatives.

2. Effective Handling of Imbalanced Data: The selected model should showcase its robustness in dealing with the imbalanced nature of the customer churn dataset. It should accurately identify instances of churn while adequately capturing non-churn instances.

3. Generalization Capability: The developed model should demonstrate the ability to generalize well to unseen data, ensuring its reliability and applicability in real-world scenarios.

4. Actionable Insights: The model's predictions should yield actionable insights that empower the telecom company to implement strategic retention measures. These insights should be clear, interpretable, and directly translatable into actionable business strategies.

5. Scalability and Integration: The solution should be scalable and seamlessly integrable into the company's operational framework. It should accommodate increasing data volumes as the customer base expands and facilitate the model's practical implementation.

6. Comprehensive Communication: The project's outcomes, methodologies, and recommendations must be communicated effectively to technical and non-technical stakeholders within the company. A clear understanding of the model's predictions and implications will drive informed decision-making.

7. Stakeholder Satisfaction: The satisfaction of key stakeholders, including management, marketing, and customer service teams, will contribute to the project's success. The model's predictions should align with their domain knowledge and complement their efforts.

8. Ongoing Performance Assessment: The project's impact will be continually evaluated to assess the effectiveness of the implemented strategies based on the model's predictions. This iterative assessment will facilitate continuous enhancements to both the model and the strategies.

Adhering to these criteria will empower the Iranian telecom company to leverage advanced analytics for customer churn prediction, enhance customer retention initiatives, and foster business growth.

2. Data

This dataset is randomly collected from an Iranian telecom companyâ€™s database over a period of 12 months. A total of 3150 rows of data, each representing a customer, bear information for 14 columns.

All of the attributes except for attribute churn is the aggregated data of the first 9 months. The churn labels are the state of the customers at the end of 12 months. The three months is the designated planning gap.

These are the 14 columns in the dataset:

1. Anonymous Customer ID (I removed this column for the train test split as it is not useful for machine learning).
2. Call Failures: number of call failures
3. Complains: binary (0: No complaint, 1: complaint)
4. Subscription Length: total months of subscription

5. Charge Amount: Ordinal attribute (0: lowest amount, 9: highest amount)
6. Seconds of Use: total seconds of calls
7. Frequency of use: total number of calls
8. Frequency of SMS: total number of text messages
9. Distinct Called Numbers: total number of distinct phone calls
10. Age Group: ordinal attribute (1: younger age, 5: older age)
11. Tariff Plan: binary (1: Pay as you go, 2: contractual)
12. Status: binary (1: active, 2: non-active)
13. Churn: binary (1: churn, 0: non-churn) - Class label

3. Data Wrangling and Cleaning

During this stage of the project I discovered the data columns in depth. I found the min, max, mean, standard deviation of each column. I found the value counts, unique values for a few columns. Most importantly we found that we have an imbalanced target variable, Customer Churn. More of the customers did not churn than did by a big margin. 2655 did not churn compared to 495 who did. This is good for the company but we need to account for this when we do the train/test split for our four models.

Fortunately we did not have any NaNs in the dataset. Typically this is an important aspect of a data science project. What to do with the null values is very important. Whether I should make them into the median, mean, 0 values, etc. is what we have to determine to make the data ready for machine learning to get an accurate prediction of our target variable.

During the Data Wrangling stage we found:

1. We have a wide range of call failures for each customer due to the high standard deviation of 7.263.

2. We have more no complaints than complaints due to the mean being low at .0765.

3. The average age of a customer was 30.99 and the customer value range was from 0-2165.28.

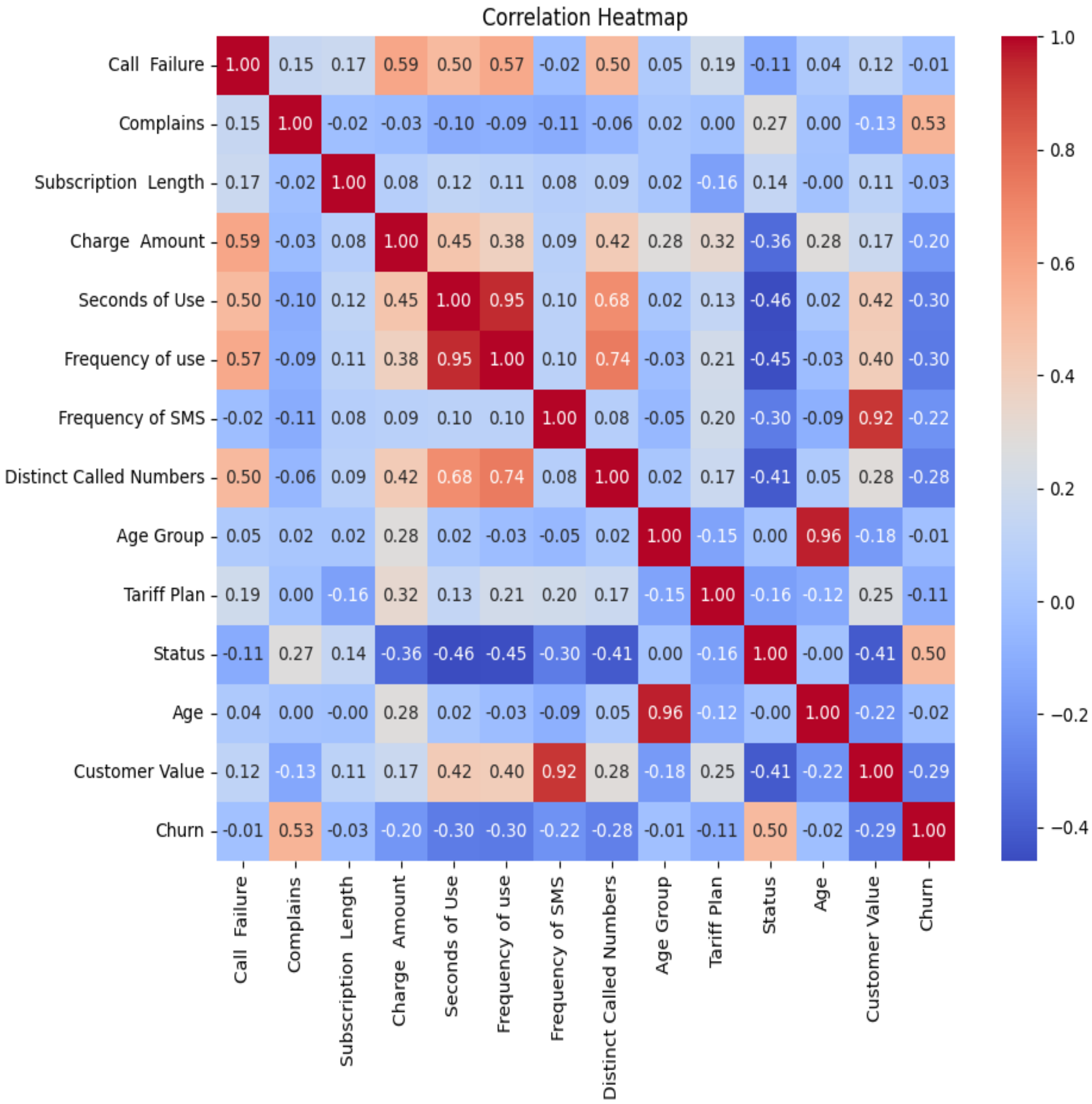
4. On average the customers used text messages more than calls.

5. We have an unbalanced target variable of Churn where we have many more customers who did not churn, 2655 compared to 495 who actually did churn.

4. Exploratory Data Analysis

Here is some interesting findings during the EDA:

4.1 Heatmap:



	Call Failure	Complains	Subscription Length	Charge Amount	Seconds of Use	Frequency of use	Frequency of SMS	Distinct Called Numbers	Age Group	Tariff Plan	Status	Age	Customer Value	Churn
Call Failure	1.00	0.15	0.17	0.59	0.50	0.57	-0.02	0.50	0.05	0.19	-0.11	0.04	0.12	-0.01
Complains	0.15	1.00	-0.02	-0.03	-0.10	-0.09	-0.11	-0.06	0.02	0.00	0.27	0.00	-0.13	0.53
Subscription Length	0.17	-0.02	1.00	0.08	0.12	0.11	0.08	0.09	0.02	-0.16	0.14	-0.00	0.11	-0.03
Charge Amount	0.59	-0.03	0.08	1.00	0.45	0.38	0.09	0.42	0.28	0.32	-0.36	0.28	0.17	-0.20
Seconds of Use	0.50	-0.10	0.12	0.45	1.00	0.95	0.10	0.68	0.02	0.13	-0.46	0.02	0.42	-0.30
Frequency of use	0.57	-0.09	0.11	0.38	0.95	1.00	0.10	0.74	-0.03	0.21	-0.45	-0.03	0.40	-0.30
Frequency of SMS	-0.02	-0.11	0.08	0.09	0.10	0.10	1.00	0.08	-0.05	0.20	-0.30	-0.09	0.92	-0.22
Distinct Called Numbers	0.50	-0.06	0.09	0.42	0.68	0.74	0.08	1.00	0.02	0.17	-0.41	0.05	0.28	-0.28
Age Group	0.05	0.02	0.02	0.28	0.02	-0.03	-0.05	0.02	1.00	-0.15	0.00	0.96	-0.18	-0.01
Tariff Plan	0.19	0.00	-0.16	0.32	0.13	0.21	0.20	0.17	-0.15	1.00	-0.16	-0.12	0.25	-0.11
Status	-0.11	0.27	0.14	-0.36	-0.46	-0.45	-0.30	-0.41	0.00	-0.16	1.00	-0.00	-0.41	0.50
Age	0.04	0.00	-0.00	0.28	0.02	-0.03	-0.09	0.05	0.96	-0.12	-0.00	1.00	-0.22	-0.02
Customer Value	0.12	-0.13	0.11	0.17	0.42	0.40	0.92	0.28	-0.18	0.25	-0.41	-0.22	1.00	-0.29
Churn	-0.01	0.53	-0.03	-0.20	-0.30	-0.30	-0.22	-0.28	-0.01	-0.11	0.50	-0.02	-0.29	1.00

High Correlation:

1. **.96, Age & Age Group**. five different groups and five ages in this database.

2. **.95, Seconds of Use & Frequency of use**. This makes sense as more calls and seconds of calls go hand in hand.

3. **.92, Customer Value & Frequency of SMS**. This was the most interesting correlation. As the company values customers who use text messages.

Low Correlation:

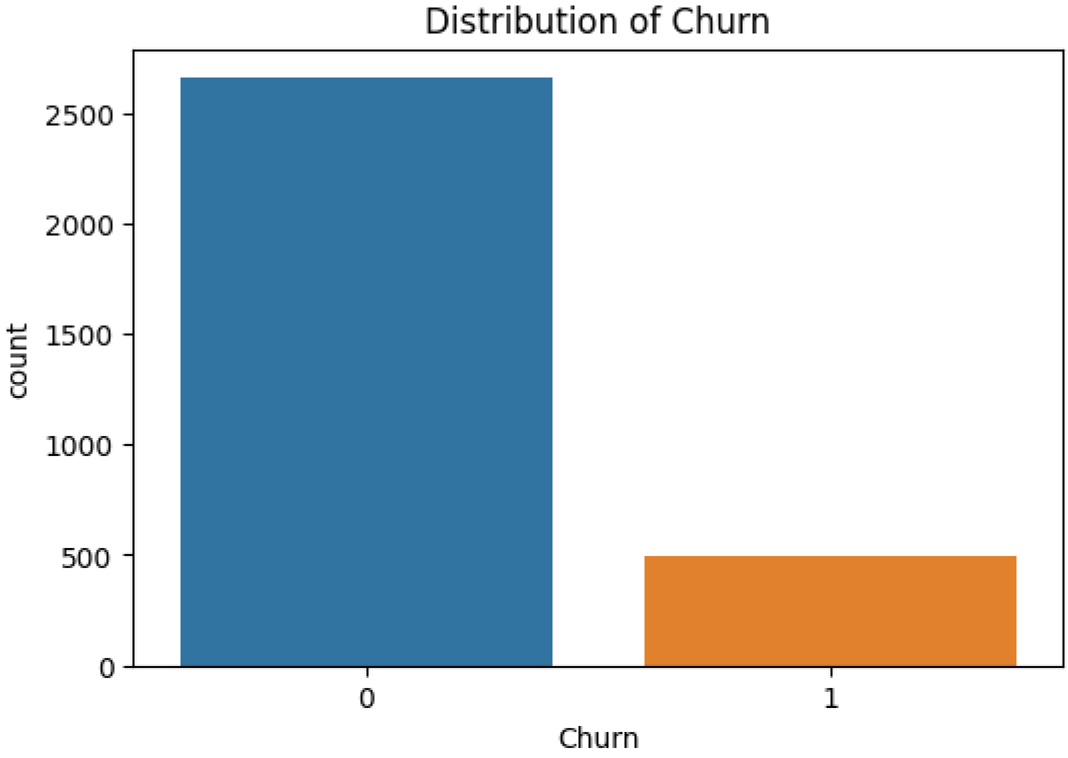
1. **.46, Status & Seconds of use**. Obviously there is no correlation if the status of a customer is inactive, (2 for inactive and 1 for active) and seconds of telephone calls.

2. **.45, Status & Frequency of Use**. This makes sense just like the previous correlation.

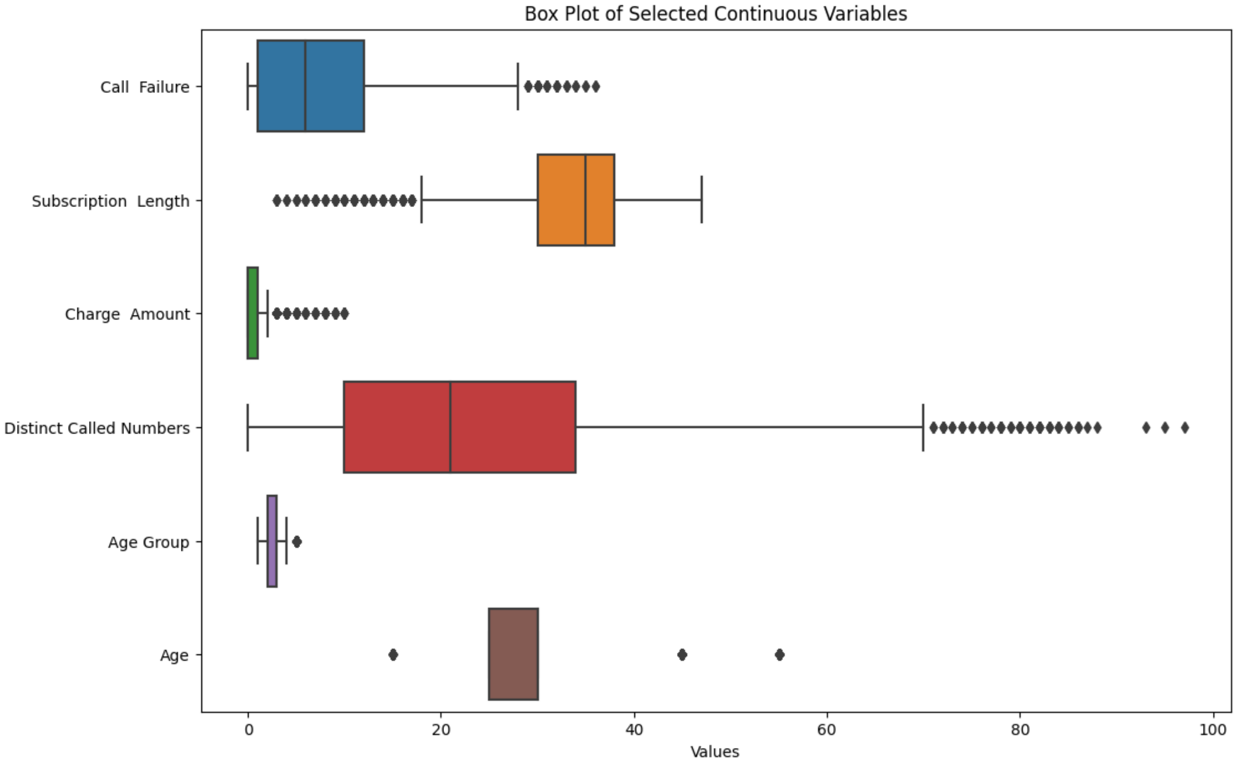
3. **.41, Status & Distinct Called Numbers**. Customers cannot dial more numbers if they are inactive.

4.2 Distribution of Churn

Then I created a graph of the unbalanced target variable, Churn. This showed us just how big of a difference we had in the customers who did and did not churn.

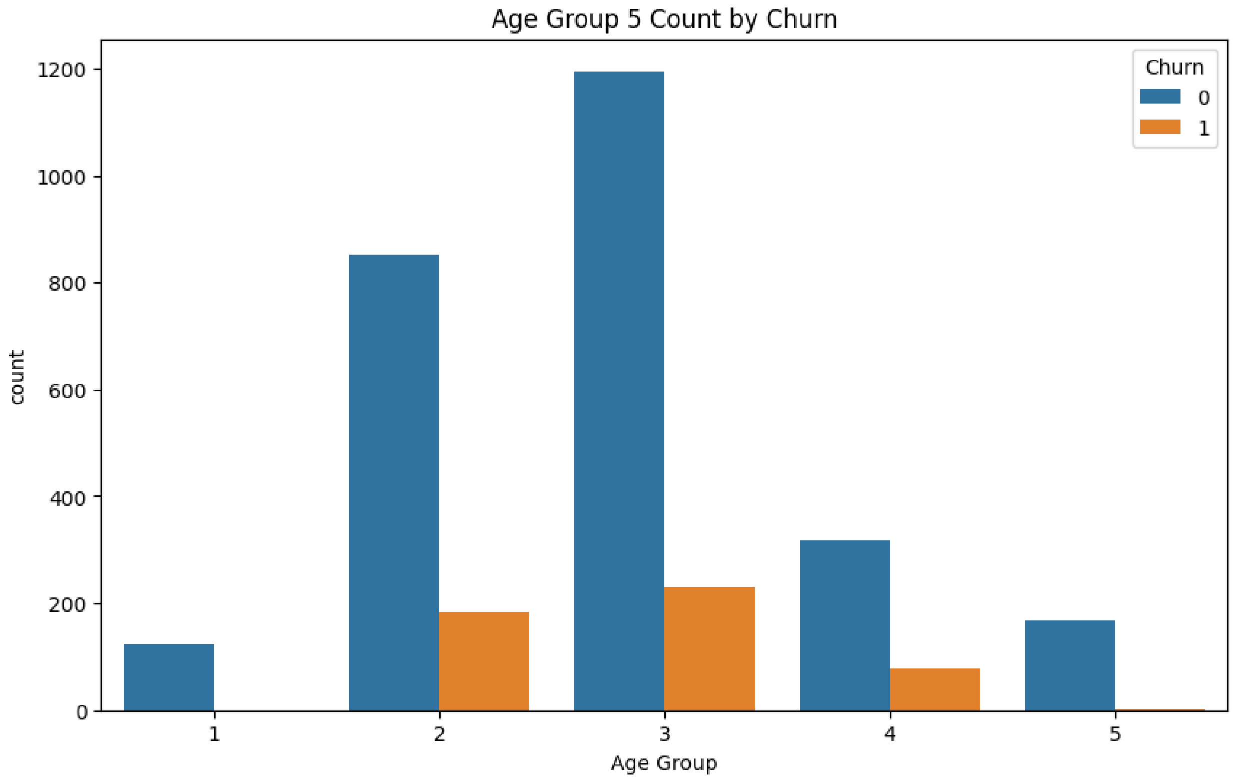


4.3 Box plot of Dependent Variables:



1. Many low outliers for subscription columns that are below the 25% percentile.
2. As well as many high outliers for the distinct called (phone) numbers column that are above the 75th percentile.
3. The median for call failure is below 10 for each customer.

4.4 Age Group Count by Churn:



Here no customers in the youngest group canceled their membership. In the oldest group bracket only a small amount did not churn. Very helpful for the company as they can target people in the other three groups to retain them. Also they could find out why customers in age groups 2-4 discontinued their services and see if they can fix the issue in a manner that is good for the business. The company needs to find out why group three had the highest churn rate.

5. Modeling Phase

Initially I did a train test split, (80/20) on the data to see how accurate each of the four models performed on the data. The four models I tried were Logistic regression, Random Forest, and Support Vector Machine, Gradient Boosting.

Gradient Boosting emerged as the most robust and effective solution for predicting customer churn within the context of an Iranian telecom company. When compared to other models Gradient Boosting consistently demonstrated superior performance in terms of both ROC-AUC and F1 scores. This model effectively balanced the trade-off between precision and recall, offering a comprehensive view of predictive accuracy on the imbalanced churn dataset.

Gradient Boosting achieved the highest accuracy of 94.60%, indicating that it correctly predicted customer churn in a significant proportion of cases. The model exhibited a high precision of 88.00%, signifying that when it predicted a customer would churn, it was accurate about 88.00% of the time. Additionally, the recall score of 80.00% indicated that the model could correctly identify 80.00% of actual churned customers.

GB’s F1-Score of 83.81% was the highest among all models, implying an optimal balance between precision and recall. This is crucial in an imbalanced dataset like this, where both false positives and false negatives need to be minimized.

The ROC-AUC Score for Gradient Boosting was 88.85%.

Its ability to discern between positive and negative classes, coupled with its strong handling of imbalanced data, positioned Gradient Boosting as the optimal choice for delivering actionable insights to the telecom company's retention strategies.

The model's reliability, interpretability, and consistent generalization to unseen data underscore its suitability for practical implementation, making it a valuable asset in addressing the challenge of customer churn.

6. Recommendations

After exploring our model and data I recommend these ideas:

1. **Tailored Retention Strategies**: Leverage the predictive power of the Gradient Boosting model to identify high-risk customers who are more likely to churn. Devise targeted retention strategies, such as personalized offers, discounts, or enhanced customer support, to proactively engage and retain these customers.

2. **Proactive Customer Engagement**: Utilize the model's ability to forecast churn probabilities to initiate proactive engagement with customers displaying early signs of disengagement. Timely intervention, through targeted communication and resolution of issues, can significantly improve customer satisfaction and reduce churn rates.

3. **Segmented Marketing Campaigns**: Leverage the model's insights to segment the customer base into distinct groups based on their churn propensity. Craft marketing campaigns tailored to the unique characteristics and needs of each segment, optimizing the allocation of resources and ensuring a more resonant and impactful message. Work on the 2-4 customer age group to reduce the higher amount of churn for those particular customer groups.

4. **Enhanced Customer Experience**: Focus on improving the overall customer experience by identifying pain points or areas of dissatisfaction highlighted by the model's feature importance analysis. Implement measures to address these concerns, such as streamlining processes, enhancing service quality, or introducing new features that align with customer preferences.

5. **Churn Prediction Monitoring**: Establish an ongoing monitoring system that regularly evaluates the performance of the Gradient Boosting model and recalibrates it as needed. As customer behavior and market dynamics

evolve, staying vigilant about model accuracy and recalibrating its parameters ensures its continued efficacy.

6. **Data Enrichment and Feature Engineeri**ng: Continuously enrich the dataset with additional relevant data sources that could provide deeper insights into customer behavior and preferences. Invest in feature engineering techniques to extract more meaningful and predictive variables, thereby enhancing the model's discriminatory power.

7. **Employee Training and Support**: Equip customer-facing teams with the knowledge and tools to utilize churn predictions effectively. Provide training on interpreting model outputs and translate them into actionable steps during customer interactions.

8. **Long-Term Customer Value:** Shift the focus beyond short-term retention and consider the long-term value of customers. The model's predictions can aid in identifying not only potential churners but also high-value customers deserving of special attention and rewards.

By aligning strategies with the predictions and recommendations provided by the Gradient Boosting model, the telecom company can optimize its customer retention initiatives, foster customer loyalty, and ultimately achieve a competitive edge in a dynamic market landscape.