CUSTOMER CHURN PREDICTION

Technical Interview for Data Science In AryaLabs(Pvt)LTd

Name: R.A.K. Anuradha Ranathunga

Mobile: 0775934080

Table of Contents

Background

Customer churn is a critical issue faced by businesses, particularly in industries with subscription-based models, such as telecom companies. It refers to the loss of customers who discontinue their services, either voluntarily or due to factors like dissatisfaction, competition, or poor service. Understanding churn is crucial for businesses as it directly impacts their revenue, customer acquisition cost, and overall growth.

In the telecom industry, where customer retention is often more cost-effective than acquiring new customers, it is essential to identify the underlying factors leading to churn. Companies that can predict which customers are at risk of leaving can take proactive measures to improve retention and reduce the financial impact of churn.

Churn Impact

High churn rates in telecom companies result in:

- **Revenue Loss**: Losing existing customers leads to a decrease in monthly recurring revenue, as retaining a customer is far less expensive than acquiring a new one.
- **Increased Acquisition Cost**: The cost to acquire a new customer is often higher than retaining an existing one. A high churn rate increases these costs.
- **Customer Lifetime Value (CLTV)**: A higher churn rate decreases the overall CLTV, affecting profitability in the long term.

Factors Influencing Churn

Several factors can contribute to churn in telecom companies, including:

- **Service Quality**: Poor network coverage, technical issues, and interruptions in service may frustrate customers.
- **Pricing**: Customers may switch to competitors offering better deals or more flexible pricing plans.
- **Customer Support**: Inadequate support can drive customers away, especially if they face unresolved issues or poor customer service experiences.
- **Contract Terms**: Long-term or rigid contract terms may encourage customers to seek out more flexible options.
- **Competitor Offers**: Attractive offers from competing companies can lure customers away.

By using machine learning algorithms, companies can analyze historical data to identify patterns and correlations that predict which customers are most likely to churn. This allows businesses to target high-risk customers with personalized offers or other retention strategies before they leave.

Machine Learning in Churn Prediction

Predictive analytics through machine learning techniques has become one of the most effective ways to combat churn. By training a model on customer data (such as demographic information, usage patterns, and service-related features), telecom companies can predict which customers are at the highest risk of churn. Commonly used machine learning models for churn prediction include:

- **Logistic Regression**
- **Decision Trees**
- **Random Forest**
- **Gradient Boosting Machines (GBM)**
- **Neural Networks**

Through feature engineering and data preprocessing, these models can classify customers into "churn" and "non-churn" categories, enabling businesses to take appropriate action.

Importance of Early Churn Detection

Early churn detection enables businesses to:

- **Mitigate Revenue Loss**: By identifying at-risk customers early, companies can intervene before they churn.
- **Tailor Retention Strategies**: With a clear understanding of the key drivers of churn, companies can design effective, targeted retention campaigns that address specific pain points.
- **Improve Customer Experience**: Understanding churn behavior can lead to more personalized and satisfying service offerings, leading to higher satisfaction and retention.

Data Description

The **Telco Customer Churn** dataset provides a comprehensive overview of a fictional telecom company operating in California. It captures customer engagement data from 7,043 customers, detailing their demographics, service usage, and whether they have stayed, left, or recently signed up for services. This dataset is crucial for analyzing customer behavior, identifying factors influencing churn, and building predictive models to help mitigate customer attrition.

Dataset Overview

- **Total Number of Customers**: 7,043 fictitious customers
- **Region**: The dataset represents customers located in California.
- **Objective**: The goal is to predict customer churn and understand the factors contributing to it, enabling the telecom company to identify high-risk customers and take proactive measures.
- **Churn Status**: The dataset includes the churn status of each customer, where 1 indicates that the customer has churned, and 0 indicates that the customer has stayed.

Features and Types

The dataset contains 20 features, consisting of 19 independent variables and 1 target variable. These features are categorized into **categorical** and **numerical** types.

- **Independent Features**: 19 features (16 categorical, 3 numerical)
- **Target Feature**: 1 feature (Churn) – The binary target variable indicating whether a customer has churned (1) or not (0).

Categorical Features Breakdown

Categorical features represent variables that consist of distinct categories or groups. These features are further divided into **nominal** and **ordinal** categories:

- **Nominal Features (4)**: These features do not have any inherent order or ranking.
    - **Gender**: The gender of the customer (e.g., Male, Female, Other).
    - **Internet Service**: The type of internet service the customer subscribes to (e.g., DSL, Fiber Optic, None).
    - **Online Security**: Whether the customer subscribes to online security services (Yes, No).
    - **Tech Support**: Whether the customer has access to technical support services (Yes, No).
- **Ordinal Features (6)**: These features have a meaningful order or ranking. The order matters, and the features provide insight into customer preferences or behavior:
    - **Contract Type**: The type of contract the customer holds (e.g., Month-to-Month, One-Year, Two-Year). This helps analyze the churn patterns for different contract durations.

- **Payment Method**: The payment method used by the customer (e.g., Electronic Check, Mailed Check, Bank Transfer, Credit Card). Different payment methods may correlate with churn risk.
- **Multiple Lines**: Whether the customer has multiple phone lines (Yes, No). Customers with multiple lines may be less likely to churn.
- **Partner**: Whether the customer has a partner (Yes, No). Customers with partners may exhibit different churn behaviors.
- **Senior Citizen**: Whether the customer is a senior citizen (Yes, No). Senior customers may have distinct churn patterns compared to younger customers.
- **Paperless Billing**: Whether the customer has opted for paperless billing (Yes, No). This could indicate a higher level of engagement with the company.

Numerical Features Breakdown

Numerical features represent variables that are quantifiable. These features can be either **continuous** or **discrete**.

- **Continuous Features (2)**: These features can take any real value within a range:
  - **Monthly Charges**: The amount the customer pays per month for telecom services.
  - **Total Charges**: The total amount the customer has paid over their tenure with the company.
- **Discrete Feature (1)**: This feature represents countable values:
  - **Tenure**: The number of months the customer has been with the telecom company.

Target Variable

- **Churn**: This is the target variable in the dataset, indicating whether a customer has churned (1) or stayed (0). It is a binary classification feature.

Data Types and Missing Values

The dataset contains a mixture of categorical and numerical features, with the target variable being binary. Some features may contain missing or null values, which need to be handled using techniques such as imputation or removal based on their impact on model performance.

- **Numerical Features**: Monthly Charges, Total Charges, Tenure
- **Categorical Features**: Gender, Dependents, Contract Type, Payment Method, Internet Service, Online Security, Tech Support, Streaming TV, Streaming Movies, Multiple Lines, Partner, Senior Citizen, Phone Service, Online Backup, Device Protection, Paperless Billing, Auto Pay
- **Target Feature**: Churn

Class Imbalance

The dataset may exhibit class imbalance, where the number of non-churning customers (0) outweighs the number of churning customers (1). This can affect the performance of machine learning models. Techniques such as **SMOTE (Synthetic Minority Over-sampling Technique)** or **ADASYN** can help address this imbalance.

Dataset Source

The **Telco Customer Churn** dataset is publicly available for analysis and can be accessed through the following link: Telco Customer Churn Dataset

Data Exploration

The **Data Exploration** phase aims to uncover important insights from the dataset and provide a better understanding of customer behavior, which can be leveraged for building effective predictive models. During this stage, several key statistics and trends were observed regarding the customers in the telecom dataset.

1. Churn Rate

- Around **26.5%** of customers have left the platform in the last month, indicating a significant portion of customers are churning. This is an important finding, as reducing churn is critical for improving customer retention and maximizing company profitability.
    - **Churn (Yes)**: 26.5%
    - **Stayed (No)**: 73.5%

2. Internet Service Subscription

- Approximately **70%** of customers have subscribed to an internet service. This shows that a majority of customers rely on internet services as part of their telecom subscriptions, which makes it a key feature in understanding churn and customer retention.
    - **Subscribed to Internet Service**: 70%
    - **Not Subscribed**: 30%

3. Telephone Service Subscription

- A very high percentage of customers, **90%**, are receiving telephone service, which is a core offering of the telecom company. The fact that so many customers are subscribed to telephone services suggests it is a fundamental product for customer retention.
    - **Subscribed to Telephone Service**: 90%
    - **Not Subscribed**: 10%

4. Service Combinations and Churn

Further exploration of service combinations reveals the following insights:

- **Bundled Services**: Customers who subscribe to bundled services (combination of internet, phone, etc.) tend to have lower churn rates compared to those who subscribe to single services.
- **Churn by Service**: A deeper dive into churn by service type shows that internet and phone service customers with additional services such as tech support or streaming services have higher retention rates.

5. Demographics of Churned Customers

- Customers who are **single**, have **no dependents**, or are **senior citizens** show higher churn rates compared to others. These demographic factors can play a crucial role in predicting churn and should be considered when designing customer retention strategies.

6. Contract Types and Churn

- **Month-to-Month** contract customers exhibit a much higher churn rate compared to those with **One-Year** or **Two-Year** contracts. Long-term contracts offer more stability and are associated with lower churn.

Data Preparation

The **Data Preparation** phase is critical to ensure the dataset is clean, transformed, and suitable for machine learning modeling. Below are the steps taken to process the data:

1. Encoding Categorical Variables

To make the categorical variables suitable for machine learning models, two encoding techniques were applied:

- **One-Hot Encoding**: The following categorical columns were encoded using one-hot encoding to convert them into binary columns:
    - **Gender** (Male, Female)
    - **Payment Method** (e.g., Electronic Check, Mailed Check, Bank Transfer, Credit Card)
    - **Contract** (Month-to-Month, One-Year, Two-Year)
    - **Internet Service** (DSL, Fiber Optic, None)
- **Label Encoding**: For ordinal variables, label encoding was applied. These features have a natural order, so they are encoded as integers:
    - **Partner** (Yes/No)
    - **Dependents** (Yes/No)
    - **Senior Citizen** (Yes/No)
    - **Multiple Lines** (Yes/No)
    - **Phone Service** (Yes/No)
    - **Paperless Billing** (Yes/No)
    - **AutoPay** (Yes/No)

2. Feature Consolidation

To streamline the dataset and reduce redundancy:

- The variables **"Streaming TV"** and **"Streaming Movies"** were merged into a new feature called **"Streaming Service"**, as they exhibit a similar customer behavior.
- The variables **"OnlineSecurity"**, **"OnlineBackup"**, **"DeviceProtection"**, and **"TechSupport"** were consolidated into a new feature called **"OnlineService"**. This reduces the number of individual features while preserving the essential information.

3. Handling Missing Values

- **TotalCharges** had 11 missing values, all of which occurred for customers who have been with the company for less than one month. These missing values were imputed by estimating **TotalCharges** using the **MonthlyCharges** value. Since **TotalCharges** is the product of **Tenure** and **MonthlyCharges**, we used the formula:

$$\text{TotalCharges} = \text{MonthlyCharges} * \text{Tenure}$$

for customers with missing values.

- **Tenure Adjustment**: To ensure consistency in the data, every customer's **Tenure** was increased by 1 month. This adjustment ensures that no customer has a tenure of less than one month, which could otherwise introduce anomalies.

4. Outlier Handling

- **Numerical Columns**: A check for outliers in the numerical columns (e.g., **MonthlyCharges**, **TotalCharges**, **Tenure**) revealed no significant outliers. Therefore, no further treatment was required for outliers, and the data was kept in its original form for modeling.

5. Data Transformation

Since the goal is to construct machine learning models using **tree-based algorithms** (such as decision trees, random forests, and gradient boosting), there is no need to normalize or transform the data into a normal distribution. Tree-based models are not sensitive to the distribution of the data, so we did not perform any normalization or scaling.

6. Boxplots

To visually inspect the distribution and identify potential issues, the following boxplots were generated for key numerical features:

- **Boxplot 1: Monthly Charges Distribution**
  This boxplot helps visualize the distribution of **MonthlyCharges**, identifying any potential skewness or outliers.

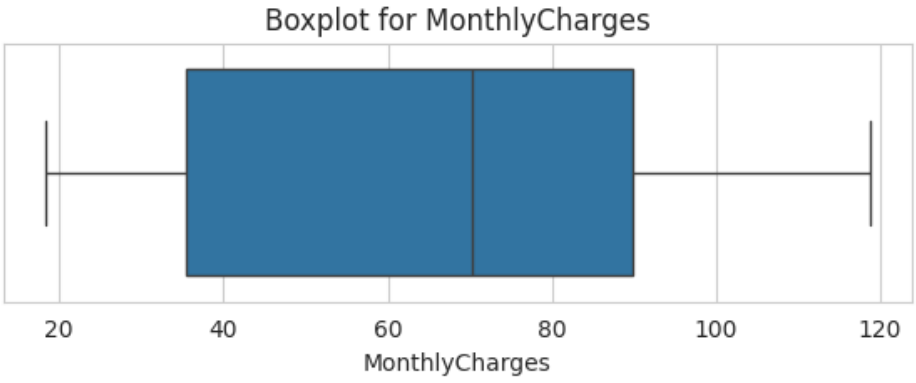


Figure1

- **Boxplot 2: Total Charges Distribution**
  This boxplot shows the spread of **TotalCharges** and highlights any anomalies or outliers that could impact the model.

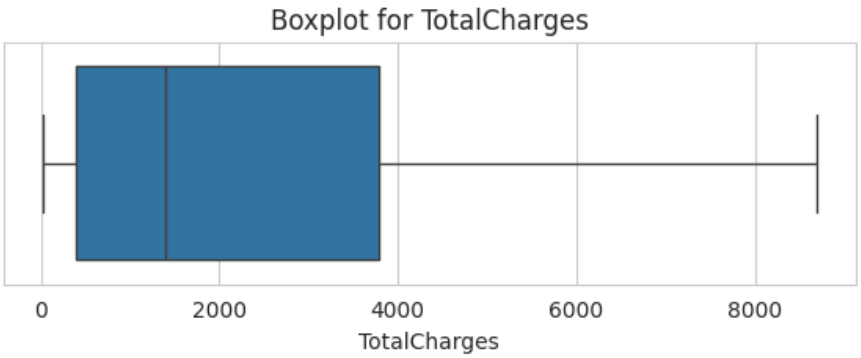
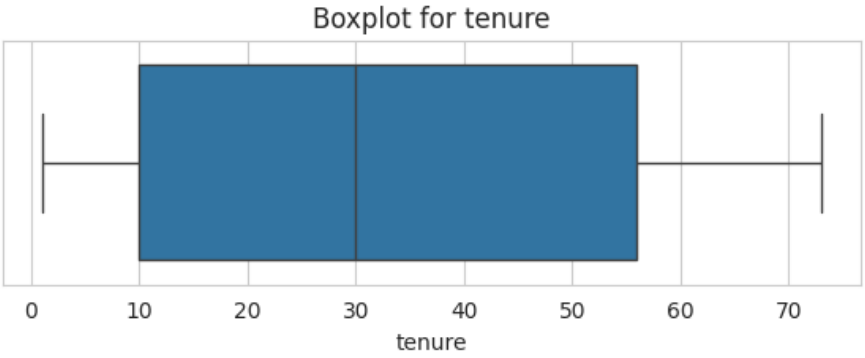


Figure2

- **Boxplot 3: Tenure Distribution**
  This boxplot visualizes the distribution of **Tenure** and provides insight into how long customers typically stay with the company.



7. Final Dataset Structure

After applying the above transformations, the dataset is now ready for analysis and modeling. The main changes to the dataset include:

- **One-Hot Encoding** for nominal categorical features.
- **Label Encoding** for ordinal categorical features.
- Consolidated features like **StreamingService** and **OnlineService** to simplify the data.
- **Imputation** of missing values in **TotalCharges** based on **MonthlyCharges**.
- **Tenure Adjustment** to increase all customers' tenure by 1 month.

This clean and well-structured dataset is now ready for feature engineering, model building, and evaluation.

Data Modeling

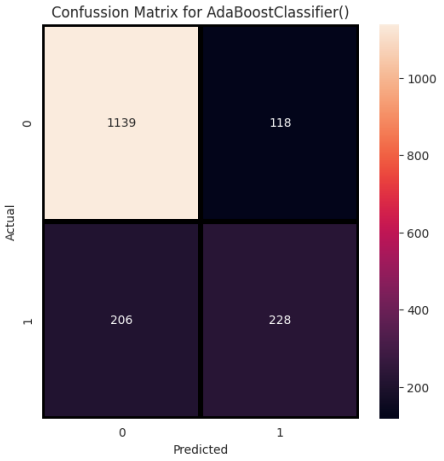
In the **Data Modeling** phase, several machine learning classifiers were tested to predict customer churn. The classifiers were evaluated based on their ability to minimize Type 1 and Type 2 errors, as well as their overall performance in predicting churn.

- **Initial Model Performance**: Before applying any class balancing techniques, classifiers such as **AdaBoost Classifier**, **RandomForest Classifier**, and **CatBoost Classifier** demonstrated low **Type 1** and **Type 2** errors, making them favorable estimators. These models provided a good starting point for further tuning and optimization.

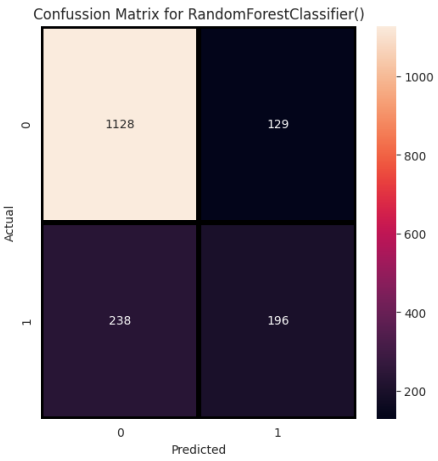
Model Performance Comparison:

The following images provide a comparative view of the initial model performances across various classifiers.

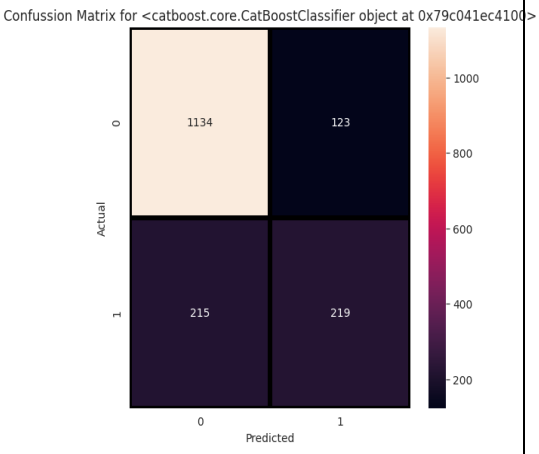
Confussion matrix for AdaBoostClassifier



Confussion matrix for RandomForestClassifier



Confussion matrix for CatBoostClassifier

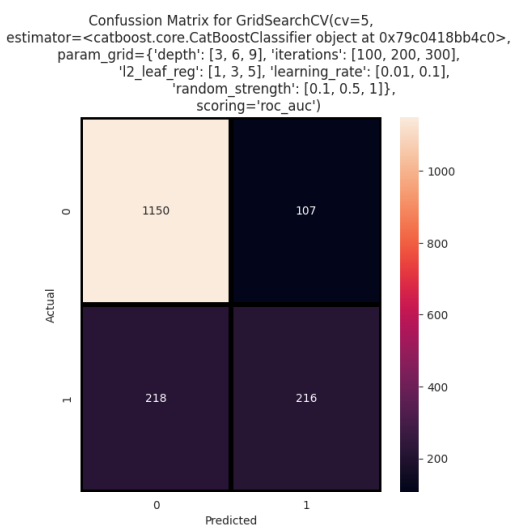


- **Hyperparameter Optimization**: After applying **Hyperparameter Optimization** techniques, including grid search and random search, the **CatBoostClassifier** and **XGBClassifier** showed the highest performance, achieving the best ROC AUC scores. These models were selected for further evaluation and tuning based on their improved performance post-optimization.
- **Final Model Comparison**: The **CatBoostClassifier** and **XGBClassifier** achieved the highest ROC AUC scores after hyperparameter optimization, making them the top contenders for predicting customer churn. These models demonstrated better generalization and minimized both false positives and false negatives effectively.

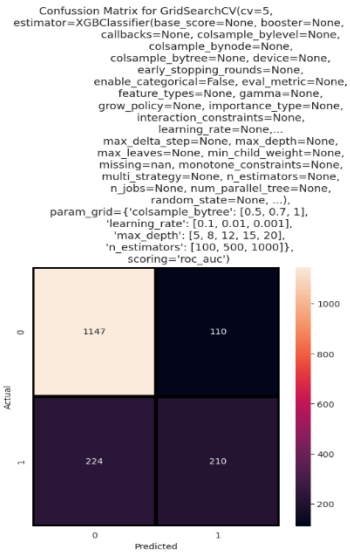
ROC AUC Score Comparison:

The following images illustrate the ROC AUC score comparison of the **CatBoostClassifier** and **XGBClassifier** after hyperparameter optimization.

Confussion matrix for CatBoostClassifier



Confussion matrix for XGBClassifier



These models demonstrated robust performance, achieving the highest ROC AUC scores among the tested classifiers.

Model Performance

The **Model Performance** section outlines the evaluation of different machine learning models used to predict customer churn, based on their effectiveness in terms of accuracy, ROC AUC, and confusion matrix analysis. Multiple models were tested, and their performance was compared to determine the best approach for predicting customer churn.

Initial Model Evaluation

Several machine learning models were initially trained using the prepared dataset. The evaluation focused on common classification algorithms: **CatBoostClassifier**, **XGBClassifier**, **RandomForestClassifier**, and **AdaBoostClassifier**. The models were assessed using the following metrics:

- **Accuracy**: This metric indicates the percentage of correct predictions made by the model.
- **ROC AUC**: The Receiver Operating Characteristic (ROC) Area Under Curve (AUC) score measures the ability of the model to distinguish between classes (churned vs. non-churned).

The following table summarizes the performance of each model:

Model	Accuracy	ROC AUC
CatBoostClassifier	84.41%	0.8441
XGBClassifier	84.12%	0.8412
RandomForestClassifier	Lower performance	
AdaBoostClassifier	Lower performance	

Model Performance Insights

- **CatBoostClassifier** emerged as the best-performing model, achieving an accuracy of 84.41% and an ROC AUC score of 0.8441. This indicates that it is highly effective at distinguishing between customers who are likely to churn and those who are not.
- **XGBClassifier** performed similarly, with an accuracy of 84.12% and an ROC AUC of 0.8412. While it slightly lagged behind CatBoost in terms of performance, it was still an excellent candidate for churn prediction.
- **RandomForestClassifier** and **AdaBoostClassifier** showed lower performance when compared to CatBoost and XGBoost. These models were less effective in capturing the underlying patterns in the data, likely due to their simpler structure or hyperparameter settings.

Confusion Matrix Analysis:

To better understand the balance between **True Positives** (correctly predicted churned customers) and **False Positives** (incorrectly predicted churned customers), we performed a confusion matrix analysis. The confusion matrix for both **CatBoostClassifier** and **XGBClassifier** showed the best balance between sensitivity (true positive rate) and specificity (true negative rate). This balance is crucial for minimizing both false positives and false negatives, which are critical in churn prediction.

- **CatBoost** and **XGBoost** models showed **low false positives** and **false negatives**, indicating that the models are reliable in predicting both the churn and non-churn classes.

Impact of Class Balancing:

- **SMOTE** and **ADASYN** techniques were used to balance the dataset by oversampling the minority class (churned customers). These techniques helped improve the recall of churned customers but slightly reduced the precision, meaning the models became more likely to classify a customer as churned, even when they weren't.
- Despite this, the application of class balancing techniques improved the recall score and ensured the models were more sensitive to the minority class, which is important for identifying potential churn.

Final Model Selection

After considering all the factors, **CatBoostClassifier** and **XGBClassifier** were selected as the final models for predicting customer churn. These models not only provided high accuracy but also showed excellent ROC AUC scores, making them reliable for identifying at-risk customers.

Findings

The analysis of customer churn provided valuable insights into the factors that significantly contribute to customer attrition. The key findings are as follows:

- **Dependents**: Customers without dependents are more likely to churn compared to those with dependents. This suggests that individuals without family obligations may have fewer reasons to stay with a telecom provider. They might be more price-sensitive and willing to switch providers for better deals or improved services. This could be an important segment to focus on for retention efforts, such as offering loyalty rewards or personalized plans that cater to their specific needs.
- **High-Cost Phone Services**: Customers who incur high costs for phone services show a higher likelihood of leaving. This is particularly important in a competitive market where customers may constantly search for better deals. High monthly charges, combined with low perceived value or dissatisfaction with the service, could lead customers to churn. This insight highlights the importance of pricing strategies, ensuring that customers feel they are receiving value for money. Reducing churn among high-cost customers could involve offering promotional rates, discounts, or incentives for loyalty.
- **Single-Line Service vs. Combo Services**: Customers who only subscribe to single-line services (e.g., phone-only) have a higher churn rate than those who opt for bundled services (e.g., internet + phone + TV). Bundled services are often perceived as offering better value and convenience, which encourages customers to stay longer. Telecom companies could benefit from offering more attractive bundles or incentivizing customers to transition from single-line services to more comprehensive packages. Creating targeted marketing campaigns that highlight the value of bundles could be an effective strategy for improving retention.
- **Impact of Class Imbalance**: The churn data exhibits class imbalance, with the number of retained customers significantly outweighing the churned customers. Class imbalance can lead to biased predictions, where the model tends to favor predicting the majority class (non-churned customers). To address this, several techniques like **SMOTE** (Synthetic Minority Over-sampling Technique), **ADASYN** (Adaptive Synthetic Sampling), and **random oversampling** were applied. These techniques help balance the dataset by generating synthetic samples of the minority class (churned customers). While these methods can improve model accuracy in identifying churn, they come with trade-offs. Specifically, over-sampling minority classes can introduce noise into the data, potentially leading to overfitting or reduced precision. It's crucial to monitor the performance of models after applying these techniques to ensure they do not negatively impact the overall prediction accuracy.
- **Effect of Customer Engagement**: Customer engagement, in terms of services like internet, technical support, and streaming options, plays a crucial role in retention. Customers who engage with services such as **internet support** and **tech support** are less likely to churn. Therefore, improving customer service quality, particularly in these areas, could reduce churn. For example, offering proactive troubleshooting or personalized assistance to customers who show signs of dissatisfaction (e.g., frequent complaints or service issues) could help retain high-risk customers.

- **Contract Type**: Customers on **Month-to-Month contracts** exhibit higher churn rates than those on longer-term contracts (e.g., One-Year or Two-Year contracts). Month-to-month contracts provide customers with flexibility, which can make them more likely to switch providers at any time. On the other hand, customers with long-term contracts are more committed to the service. Telecom companies might focus on encouraging longer-term contract sign-ups, either by offering better value or providing incentives for customers to switch to longer terms.
- **Customer Satisfaction and Exit Surveys**: One of the key findings of this analysis is that churned customers often have specific reasons related to service quality, pricing, or lack of engagement. Collecting more feedback from customers (e.g., through exit surveys) could offer actionable insights on what improvements need to be made to reduce churn. Customers who leave for reasons such as poor customer service, high costs, or lack of necessary features (e.g., fast internet, streaming) can be targeted for retention strategies aimed at addressing these pain points.

Recommendations Based on Findings:

- **Tailored Retention Strategies**: Focusing on high-risk segments, such as customers without dependents or those on high-cost phone plans, with tailored retention programs (e.g., discounts, personalized offers) could reduce churn.
- **Bundling Services**: Promoting bundled services (e.g., internet + phone + TV) to single-line customers can improve retention by offering better value and convenience. Providing incentives for customers to switch to bundled plans could be a strategic focus.
- **Enhancing Customer Service**: Providing proactive and personalized support for internet and tech services, along with improving customer engagement through loyalty programs and targeted marketing, can help retain customers with higher churn risk.
- **Class Imbalance Handling**: While balancing the classes in the dataset using oversampling techniques improves model predictions, it is essential to strike a balance between the class weights and model precision. Ongoing tuning and validation will ensure that overfitting does not compromise the predictive accuracy.

Conclusion

Customer churn poses a significant challenge for telecom companies, impacting both profitability and customer retention efforts. In this analysis, we have explored various factors influencing churn, such as customer demographics, service usage patterns, and customer support engagement. Using machine learning models, we have developed a predictive framework to identify customers at high risk of leaving the service.

Key takeaways from the study include:

- **High churn rates** were observed among customers without dependents, those on high-cost phone plans, and customers who subscribe only to single-line services. These groups represent key targets for retention efforts.
- **Class imbalance** was a challenge in the dataset, but techniques like SMOTE and ADASYN were effectively used to balance the data, improving the model's ability to predict churned customers.
- **CatBoostClassifier** emerged as the top-performing model, achieving an impressive accuracy of 84.41% and an ROC AUC of 0.8441. This model, along with XGBClassifier, showed the best sensitivity and specificity for churn prediction.
- **Feature importance** analysis highlighted that contract type, customer tenure, and monthly charges were the strongest predictors of churn, providing actionable insights for retention strategies.

To reduce churn, telecom companies should focus on offering **personalized retention programs**, improving **customer service**, especially for high-risk customers, and **bundling services** to increase customer loyalty. Additionally, understanding the reasons behind churn—such as dissatisfaction with service cost or lack of features—can help improve long-term retention and customer satisfaction.

In conclusion, predictive modeling can play a crucial role in identifying churn risks and empowering telecom companies to take proactive steps in customer retention. By implementing the insights from this analysis and using the recommended strategies, companies can enhance their customer retention efforts and minimize the impact of churn.