Group Name: 03

Vikas Kumar Shukla

Nitin Ramagiri

Vedant Kishor More

Sylvester Obeng Krampah

DS 501 – Case Study

# Introduction:

Customer churn, the loss of customers over time, is a significant issue for subscription-based businesses. It results in lost revenue and increases the cost of acquiring new customers. Our objective is to proactively identify customers at risk of churning, address their concerns, and enhance retention rates to boost overall profitability.

# Why Solving Customer Churn is Crucial:

1. **Prevent Revenue Loss**:
   1. High churn rates directly reduce revenue.
   2. Increased dependency on acquiring new customers, which is more expensive.
2. **Enhance Cost-Effectiveness**:
   1. Retaining existing customers is more economical.
   2. Offers a higher return on investment compared to acquiring new ones.
3. **Improve Customer Loyalty**:
   1. Addressing churn fosters stronger relationships with customers.
   2. Enhances brand loyalty and trust.
4. **Increase Market Competitiveness**:
   1. Reducing churn ensures a more stable customer base.
   2. Provides the company with a competitive edge in the market.

**The dataset:** comes from a telecommunications company and is designed to help predict customer churn—whether a customer decides to leave the service. It contains information about customer demographics, account details, service usage, and contract types. Some key features include the length of customer contracts (month-to-month, one-year, or two-year), monthly and total charges, and optional services like online security and tech support. The target variable, "Churn," indicates whether a customer left the service (1) or stayed (0). By analyzing this data, we can uncover the factors influencing customer loyalty and churn.

# **Business Proposition:**

Customer churn poses a critical challenge for subscription-based businesses. Our proposal is to create a predictive churn model that detects at-risk customers and enables proactive retention efforts. By utilizing machine learning techniques, including XGBoost and logistic regression, we aim to identify key drivers of churn, such as contract type, payment method, and monthly charges. This approach will help businesses lower churn rates and improve customer retention.

# **Motivation:**

Retaining customers is more cost-efficient than acquiring new ones, and reducing churn has a direct effect on a company’s profitability and long-term viability. This topic is crucial as it addresses both financial outcomes and customer satisfaction. By delving into this area, we strive to connect business goals with advanced analytics, making it a particularly compelling focus.

# **Analysis:**

For the company, it is very crucial to understand the churn factor to retain the customers. To understand this, we have analysed the data and found the following graphs:

|  |
| --- |
|  |
| Fig 1: Relationship between customer churn and **one-year contract** |

|  |
| --- |
|  |
| Fig 2: Relationship between customer churn and **two-year contract** |

The first graph examines the impact of one-year contracts on customer churn. It shows that customers without a one-year contract are more likely to leave, while those with a one-year contract are much less likely to churn. This suggests that having a one-year agreement offers some stability and helps reduce the likelihood of customers leaving the service.

The second graph highlights the effect of two-year contracts and shows an even stronger link between longer commitments and lower churn.

Customers without a two-year contract are more likely to churn, whereas those with a two-year agreement rarely leave. This indicates that longer contracts, such as two-year plans, are particularly effective in retaining customers. These longer commitments may work because of the benefits they provide or the costs associated with early termination, making them a strong strategy for improving customer retention. We have also calculated the churn rate for one year contract and two year contract from the customers to understand their behaviour regarding our services. Customers with no long-term contracts (neither one-year nor two-year) are the most likely to leave, with a churn rate of nearly 43%. On the other hand, customers with a two-year contract are the most loyal, with a churn rate of just 2.85%. Those with a one-year contract fall somewhere in between, with a churn rate of about 11%.

This breakdown shows that longer contracts, especially two-year agreements, significantly retain customers and reduce churn. Encouraging customers to sign up for longer-term contracts can be an effective strategy for improving customer loyalty.

# **Model Selection and Conjectures:**

We decided to use the logistic regression model as a classification algorithm. The aim is to predict customer churn whether a customer will leave the service ( 1) or stay (0) based on various features such as contract type, monthly charges and payment method etc.

After applying the logistic regression, we found that the model accuracy is around 79% in predicting whether the customer will churn or not. For class 0 (customer stay), we got the precision 84% and recall was 88% that shows the model identifies 88% of all actual non-churners correctly. However, in the case of class 1 ( churners, customer will leave ), the precision is 62% which shows that model is less confident in recognising the actual churners and the recall is also less around 55%.

|  |
| --- |
|  |
| Fig 3: Performance of a logistic regression classification model. |

Based on our logistic regression model, we made the first conjecture that **Customer retention is predictable based on features** where we have 79% accuracy suggesting the churn behaviour is partially predictable using customer features such as contract type, payment method, tenure, and monthly charges. This also suggests the relationship between variables and the likelihood of churn. Besides this, using the model, our second conjecture is that **non-churners are easier to predict than churners;** the recall for non-churners is 88%, which is higher than that for churners, 55%. This suggests that the model is better in recognizing the customers likely to stay compared to those who are leaving.

The above-proposed conjecture will have business implications such as encouraging customers to adopt longer-term contracts and addressing dissatisfaction early for customers with short tenure or high monthly charges.

A diagram of a confusion matrix

Description automatically generated

|  |
| --- |
|  |
| Fig 4. Confusion matrix developed by XGBoost model |

A graph showing a curve

Description automatically generated

|  |
| --- |
|  |
| Figure 5: ROC Curve for the developed model |

A graph with blue and pink dots

Description automatically generated with medium confidence

|  |
| --- |
|  |
| |  | | --- | | Figure 6: SHAP value for the features contribution in the model |   The next model used was XGBClassifier (XGBoost Classifier), which is an implementation of the gradient boosting algorithm from the XGBoost library. Specifically, the model is initialized with parameters such as random\_state=42, use\_label\_encoder=False, and eval\_metric='logloss'.   1. Data Preparation:    * The dataset is loaded from a CSV file.    * Missing values are handled by replacing empty strings with NaN and dropping rows with missing data.    * The 'TotalCharges' column is converted to numeric, and any resulting NaN values are dropped. 2. Feature Engineering:    * Categorical variables are encoded using LabelEncoder, excluding the 'customerID' column.    * Features (X) and target variable (y) are separated.    * Numerical features are standardized using StandardScaler. 3. Train-Test Split:    * The data is split into training (80%) and testing (20%) sets. 4. Class Imbalance Handling:    * SMOTE (Synthetic Minority Over-sampling Technique) is applied to the training data to address class imbalance. 5. Model Training:    * An XGBoost classifier is initialized with the following parameters:scale\_pos\_weight: Set to balance class weightsrandom\_state: Set for reproducibilityeval\_metric: Set to 'logloss' for binary classification    * The model is trained on the resampled training data.   The XGBoost classifier model for customer churn prediction demonstrates strong performance:   1. Accuracy: 85.94%, indicating the model correctly predicts churn for about 86% of customers. 2. Balanced performance: Both precision and recall are 0.86 for both classes (churned and non-churned customers), suggesting the model performs equally well for both outcomes. 3. AUC-ROC Score: 0.9409, which is excellent. This indicates the model's high ability to distinguish between churned and non-churned customers. 4. Cross-validation: The model shows consistent performance across different subsets of the data, with ROC AUC scores ranging from 0.9196 to 0.9419. 5. Mean ROC AUC: 0.934 (±0.016), demonstrating robust and stable performance across different data splits.   Overall, the model exhibits strong and consistent predictive power for customer churn, with balanced performance across classes and high discriminative ability. |

# **Business Insights and Conjecture for Decision Makers:**

The results provide meaningful insights into customer churn, enabling targeted strategies to boost retention, optimize resources, and drive revenue growth. The XGBoost model pinpoints at-risk customers, helping businesses focus on high-value retention efforts like loyalty programs or personalized offers. For instance, customers with high monthly charges or long tenure but minimal engagement with services like "TechSupport" could benefit from tailored campaigns. Retaining customers is far cheaper than acquiring new ones, and even reducing churn by 10% among 374 high-risk customers could add $37,400 annually. Additionally, the model highlights key churn drivers such as contract type or payment method, guiding personalized marketing strategies like offering discounted annual plans to month-to-month users. While false positives and negatives present opportunities for refinement, enhancing prediction accuracy through advanced techniques could unlock significant cost savings and solidify competitive advantages in customer retention.

# Investment Justification to Top Executives:

This analysis uses advanced machine learning, specifically the XGBoost model, to turn customer data into actionable insights, enabling proactive churn prevention. By focusing on targeted retention strategies, businesses can reduce churn-related costs, enhance marketing efficiency, and avoid wasteful spending on broad campaigns. The ability to predict and address churn provides a competitive edge by fostering customer loyalty in a crowded market. Additionally, the model’s scalability across various customer segments and regions ensures it remains a sustainable, long-term solution for improving customer retention and driving business growth.

# Work Division and Contribution:

We adopted a structured and collaborative approach to ensure the project’s success. We worked together seamlessly, dividing responsibilities to leverage each member’s strengths and maintaining clear communication throughout.

* **Ideation:** We began by brainstorming potential topics and unanimously decided on customer churn prediction due to its significant relevance and potential impact on subscription-based businesses. This topic was both challenging and rewarding, aligning with our shared interest in applying machine learning to solve real-world problems.
* **Data Analysis:** As a team, we explored the dataset collaboratively, diving deep into customer demographics, account details, and service usage. Each member contributed to identifying key trends and patterns, such as the relationship between contract type and churn, monthly charges, and payment methods. These insights formed the basis for our conjectures and modeling strategy.
* **Model Development:** We divided the task of model creation to ensure efficiency and coverage. Vikas and Vedant focused on developing the baseline logistic regression model, while Nitin and Sylvester implemented and refined the advanced XGBoost model. Despite extensive efforts in hyperparameter tuning and trying different algorithms (including CatBoost and LightGBM), improving the model accuracy beyond 80% remained a challenge. This obstacle tested our patience and problem-solving skills, pushing us to explore creative solutions and gain a deeper understanding of the dataset.
* **Report Writing:** In the final phase, the team collaborated to consolidate our findings and draft the report. Vikas and Nitin organized the results and visualizations, while Vedant and Sylvester ensured the narrative was cohesive and aligned with project goals. We took extra care to present the insights clearly and include actionable recommendations for business stakeholders.

**Challenge Faced:** A key challenge we encountered was achieving higher accuracy with our models. While the baseline logistic regression model performed reasonably well, our XGBoost implementation required significant tuning to improve precision and recall, particularly for identifying churners. Despite experimenting with various hyperparameters and algorithms, we observed diminishing returns in accuracy. However, this experience strengthened our analytical skills and highlighted the importance of balancing predictive performance with practical business applications.

Through teamwork, determination, and consistent effort, our group successfully completed the project, turning a complex problem into actionable insights. Each member’s unique contributions were instrumental in making this project a valuable learning experience.

**References:**

Dataset link:[https://www.kaggle.com/datasets/blastchar/telco-customer-churn?resource=download](https://nam11.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.kaggle.com%2Fdatasets%2Fblastchar%2Ftelco-customer-churn%3Fresource%3Ddownload&data=05%7C02%7Cvshukla%40wpi.edu%7C51dfcaa6b89e4f58338b08dd05a16964%7C589c76f5ca1541f9884b55ec15a0672a%7C0%7C0%7C638672912996203870%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=JonhGuMA2T811RF36VG3BMD3Jk2J%2BFwQX69B5l5uUoc%3D&reserved=0)

ChatGPT was utilized to make this content. It provided high-level suggestions for improving the overall structure and clarity of the writing, including checking for consistency in verb tense and sentence structure and ensuring the flow of information throughout the report was logical and concise.