Logo, company name

Description automatically generated

**CUSTOMER CHURN**

**ISM 6353**

**Programming for Data Analytics – Python & Machine Learning**

Professor Andy Chen

Sam Fischer, Oanh Nguyen, Shely Nunez, Nicholas Sutanto

**Table of Contents**

[1. Business Insight 3](#_Toc177472044)

[1.1. Business Understanding 3](#_Toc177472045)

[1.2. Stakeholders 3](#_Toc177472046)

[2. Data Analysis 4](#_Toc177472047)

[2.1. Data Understanding 4](#_Toc177472048)

[2.2. Data Preparation 4](#_Toc177472049)

[2.3. Exploratory Data Analysis (EDA) 6](#_Toc177472050)

[Correlation between features except for the Churn rate 6](#_Toc177472051)

[Target Variable Analysis – Churn rate 6](#_Toc177472052)

[Correlation between different features and Churn rate 7](#_Toc177472053)

[Support calls and Churn rate 8](#_Toc177472054)

[Total spend and Churn rate 9](#_Toc177472055)

[Payment delay and Churn rate 9](#_Toc177472056)

[Age distribution and Churn rate 10](#_Toc177472057)

[Last interaction and Churn rate 11](#_Toc177472058)

[3. Main Analysis: 12](#_Toc177472059)

[3.1. Overview of models used for main analysis 12](#_Toc177472060)

[3.2. Logistic Regression 12](#_Toc177472061)

[3.3. Decision tree 13](#_Toc177472062)

[3.4. XGBoost 14](#_Toc177472063)

[3.5. Clustering 14](#_Toc177472064)

[4. Recommendations: 15](#_Toc177472065)

[4.1. Personalized Support 16](#_Toc177472066)

[4.2. Increase in Customer Retention 16](#_Toc177472067)

[5. Faith and Ethics Implications 16](#_Toc177472068)

[5.1. Privacy and Data Security: 17](#_Toc177472069)

[5.2. Bias and Fairness 17](#_Toc177472070)

[5.3. Christian Ethics 18](#_Toc177472071)

# Business Insight

## Business Understanding

For this project, we have decided to use a customer churn dataset to analyze. Customer churning is an ongoing problem in many industries. Customer churn is a problem because it will lead to your business losing revenue, competitors taking your market share, impact business growth, and tarnish reputation. What is customer churn? Churn refers to the number of customers who have stopped using your product or service. The average customer churn rate is surprisingly high at 30%. Even though customer churn is inevitable, it would be a problem if your business has a high/increasing customer churn rate. Companies should understand what customer churn is because it can be expensive for companies, and it can also help predict future projections. With this project, we look to find patterns and insights to what factors can lead into customer churning.

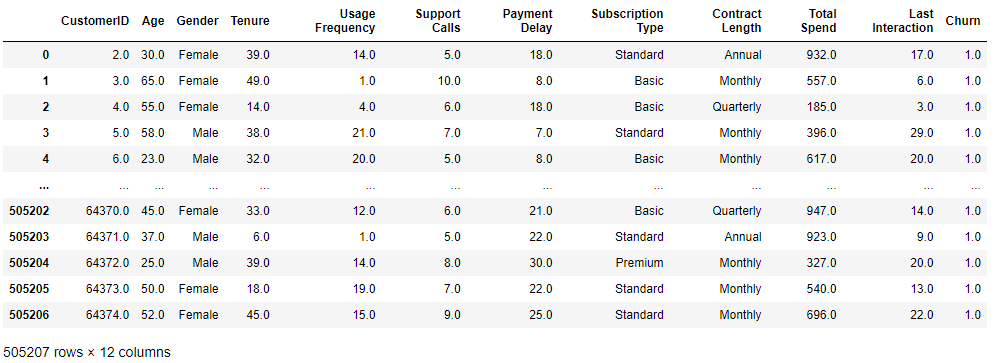
## Stakeholders

* **Executives and Management**: Executives and senior management are some stakeholders who may care about the overall health and sustainability of the business and retaining customers. High customer churn rates can impact revenue and profitability. Executives and management would be interested in insights that can help reduce churn and improve customer retention.
* **Marketing and Customer Service Team:** The marketing and customer service team are likely interested in understanding the factors that contribute to customer churn. Insights from the analysis can guide marketing strategies to attract and retain customers more effectively. By identifying patterns that lead to customer churn, the teams gain valuable insights into recurring issues and common customer concerns. This understanding allows teams to proactively address potential problems, reaching out to customers showing early signs of dissatisfaction to offer solutions and support. This serves as a way to improve overall customer satisfaction by refining communication strategies, enhancing support processes, and providing feedback for product and service enhancements.
* **Product Development Team:** Members of the product development team may be interested in customer churn analysis because the analysis provides valuable insights that directly influence the evolution of products and services. By gathering data and information about the reasons behind customer churn, the team gains a clear understanding of where the product or service may be falling short of meeting customer expectations. This information serves as a roadmap for strategic improvements, allowing the team to prioritize enhancements and introduce new features that directly address customer needs. Customer churn analysis acts as feedback for the development team, encouraging them to make data-driven decisions that can contribute to higher customer satisfaction and loyalty.

# Data Analysis

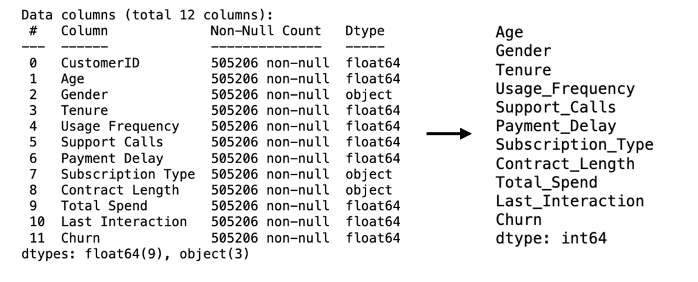
## Data Understanding

The dataset was retrieved from Kaggle. It contains two different files, one for training and one for testing. Both datasets combined have a total of over 500,000 rows and 12 columns. Each row in the dataset represents a customer. Each column represents a feature such as CustomerID, Age, Gender, Tenure, Usage Frequency, Support Calls, Payment Delay, Subscription Type, Contract Length, Total Spend, and Churn. The dataset we chose leans more towards numerical variables than categorical. Each customer has a unique CustomerID. Tenure represents how long the customer has used the product/service; Usage Frequency represents how many times the customer has used the product/service; Support Calls show the amount of times the customer has contacted customer support; Payment Delay indicates the amount of days passed since the customer has had to pay for the product/service; Subscription Type shows the type of subscription; Contract Length represents the length of the product/service contract; Total Spend indicates the total amount spent by that customer; and lastly, Churn indicates in boolean form whether the customer has stopped using the product/service.

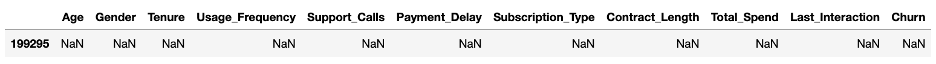
The figure below shows the dataset:

## Data Preparation

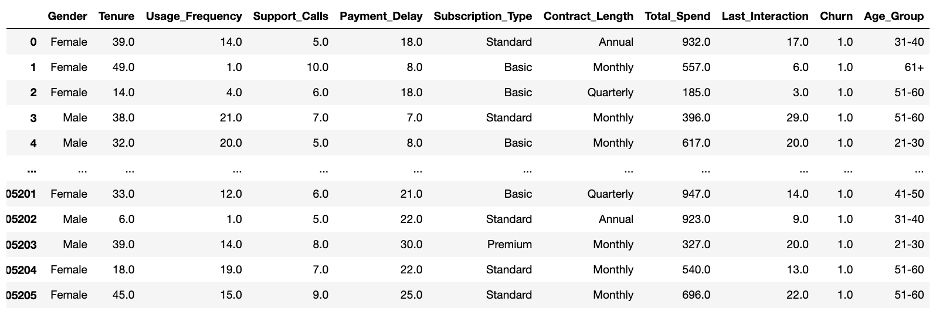
Initially, the Customer ID, which serves as a distinct numerical identifier for each customer within the organization, was eliminated. Subsequently, the vacant space that is present between variable names was substituted with an underscore character ("\_"), which facilitated the process of referencing said variable.



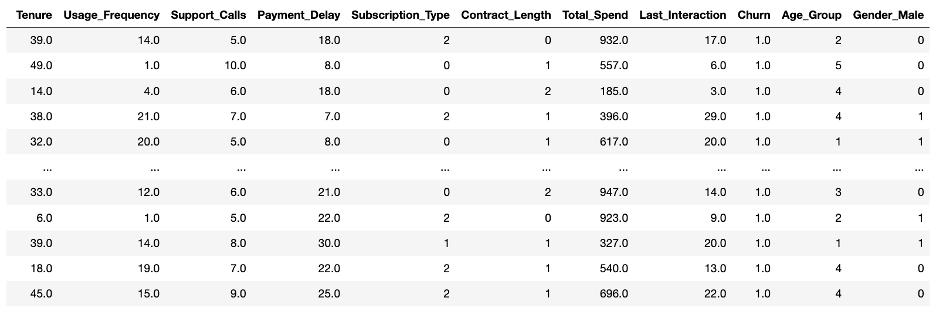
Given the cleanliness of the dataset, few preparatory measures were required, with the sole task of eliminating null values.



Furthermore, in order to investigate and examine the data, we employed a technique of converting a continuous variable into a categorical variable by dividing the age values of customers into distinct intervals of age ranges, specifically 0-20, 21-30, 31-40, 41-50, 51-60, and 61+, with the intention of facilitating the process of visual representation.



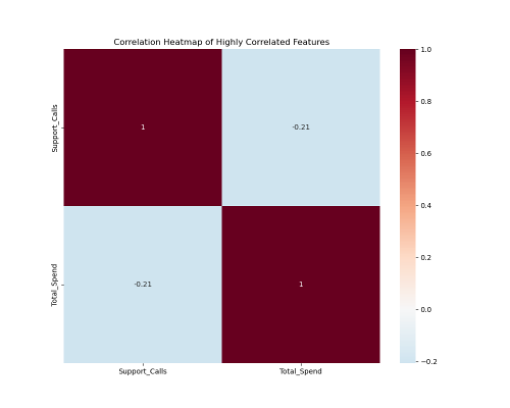
Subsequently, the string labels were transformed into numerical labels in order to facilitate the execution of the model. In addition, it was necessary to convert other variables such as Gender, Subscription\_Type, and Contract\_Length into numerical labels for the primary analysis.



## Exploratory Data Analysis (EDA)

### Correlation between features except for the Churn rate

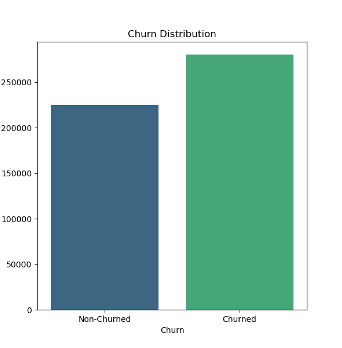
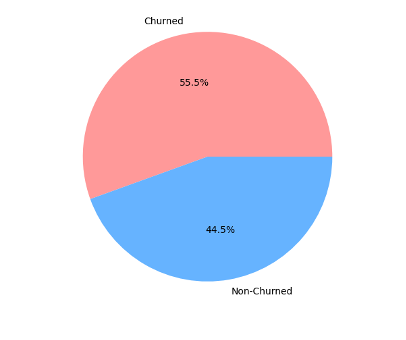
Firstly, we tested the correlation between features except for the Churn rate.



The resulting plot shows there are only 2 features (including total spend and support calls) that are highly correlated except for Churn at 0.21. Therefore, we kept all the variables of this dataset for analysis.

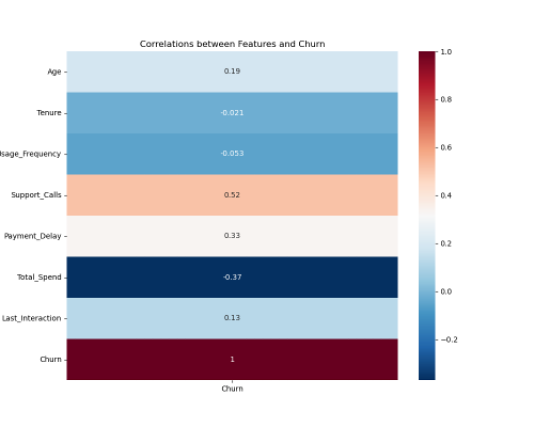
### Target Variable Analysis – Churn rate

In the first place, we analyzed the target variable which is the Churn rate. By analyzing that, we can see most customers of the shop are churning!

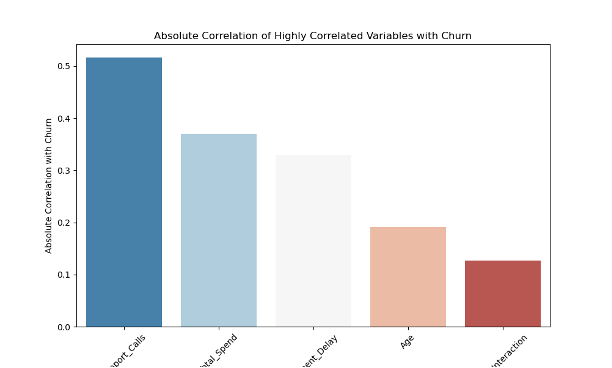
****

### Correlation between different features and Churn rate

Next, we tested to see if there were any notable correlations between different features and churn.



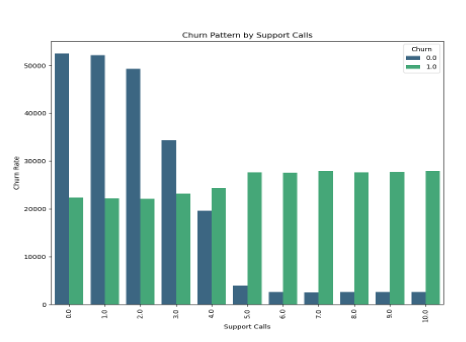
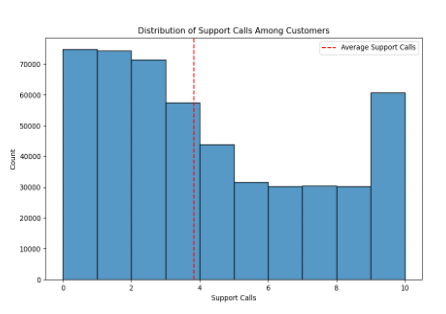
From the first plot, we realize that the support calls feature has the highest correlation with Churn.



The second plot shows the features that have the highest correlation with the target variable.

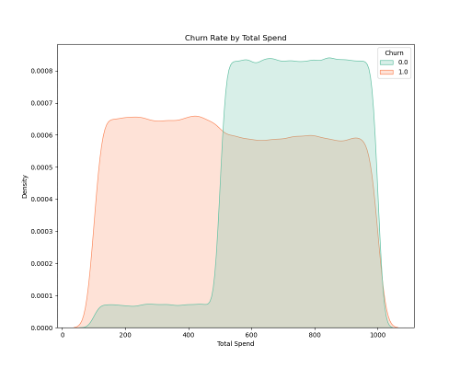
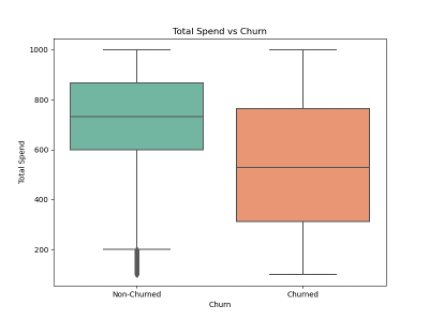
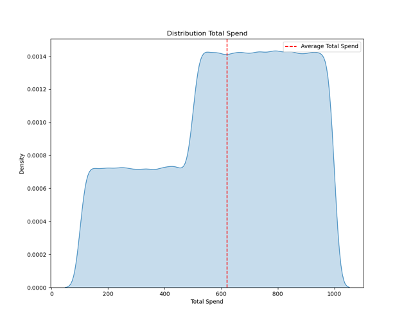
Therefore, we will do the exploratory data analysis related to these 5 important variables.

### Support calls and Churn rate



On average, clients contact the support service three times. There is a notable increase in support calls ranging from 0 to 3 instances per customer. Within this range, the calls appear to have positive outcomes, as evidenced by the churn rate of these customers. As the volume of support calls increases, there is a corresponding increase in churn rates. This relationship is particularly evident when the number of calls exceeds five, as there are no remaining customers who have not churned. This suggests that the support team's effectiveness in resolving customer issues may be limited.

### Total spend and Churn rate



On average, the expenditure per customer amounts to approximately $650. A significant proportion of consumers who exceed a spending threshold of $500 exhibit a lower likelihood of churning.

### Payment delay and Churn rate

It is normally distributed and slightly positively skewed. The churn rate exhibits a notable increase within the initial 20 months but experiences a substantial decline subsequent to a 20-month period of payment deferral.

A diagram of a distribution

Description automatically generatedA graph of a number of blue and green lines

Description automatically generated

It is normally distributed and slightly positively skewed. The churn rate exhibits a notable increase within the initial 20 months but experiences a substantial decline subsequent to a 20-month period of payment deferral.

### Age distribution and Churn rate

A graph of a number of people

Description automatically generated with medium confidenceA chart of a diagram

Description automatically generated with medium confidenceA graph of age distribution

Description automatically generated

The distribution of age follows a normal distribution, albeit with a modest positive skew, indicating a prevalence of younger customers.

A graph of different colored bars

Description automatically generated with medium confidence

Furthermore, we conducted a segmentation of the customer age variable into six distinct categories. Upon analysis, it became evident that a considerable proportion of customers fell within the 40-50 age bracket, while a smaller nevertheless noteworthy cohort was observed within the 20-30 age range. Approximately 50% of consumers within the age range of 20-30 exhibit churn behavior. However, the churn rate gradually decreases for subsequent age groups until reaching the 41-50 age group. Individuals who are beyond the age of 60 exhibit a propensity for churning.

### Last interaction and Churn rate

A graph with blue lines

Description automatically generatedA diagram of a surfboard

Description automatically generated

It was observed that the final engagement of customers has a uniform distribution but with a small elevation within the range of 0 to 15. There is no substantial difference in the churn rate based on the last engagement, indicating that the last interaction has minimal or no impact on the churn rate.

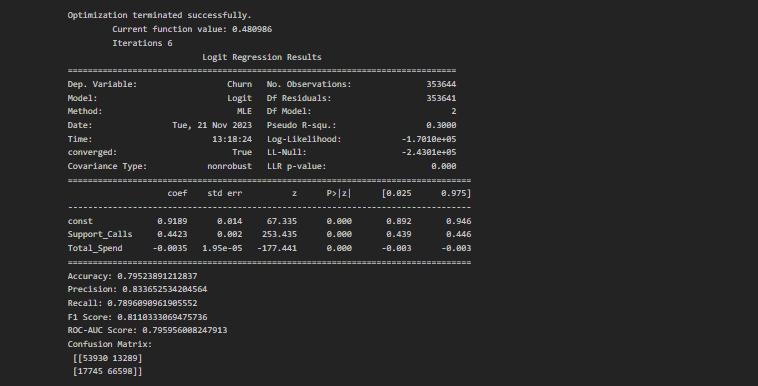
# Main Analysis:

## Overview of models used for main analysis

Our main goal was to determine which feature had the greatest impact on customer churn. In order to do this, we first ran a feature importance model to identify which variable had the highest total reduction of criterion or, in other words, the most impact on customer churn. Once we had identified the feature with the greatest impact on customer churn we used that variable as our “y” variable in a logistic regression model. After the logistic regression model was used, we implemented a decision tree to predict the possible outcomes of customer churn.

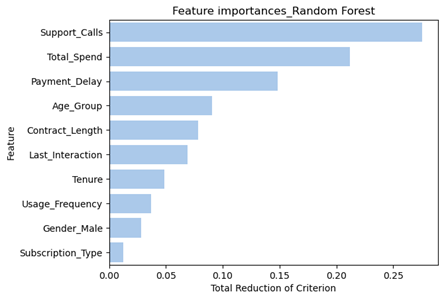
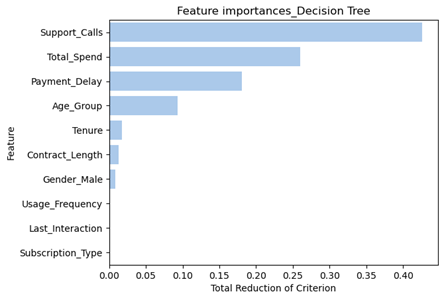
## Logistic Regression

Our logistic regression model was applied to identify the features that are impacting customers that are churning. We will be using “Support\_Calls” and “Total\_Spend” as our X and Y variables outlined in our aforementioned feature importance model in relation to customer churn. As a summary, our logistic regression model showed the following: Our support call metric had a .4423 coefficient suggesting that an increase in support calls had a near correlation to customer churn. The more support calls, the higher likelihood of the customer churning with every support call n+1. Since our P-value was practically zero, this indicates that support calls had a high significance in overall customer churn. Our total spending feature had a negative coefficient of -.0035 which implies a higher rate of spending has a negative relationship with customer churn. The more customers spend the less likely they are to churn. This was also significant as it scored a near zero P-value.

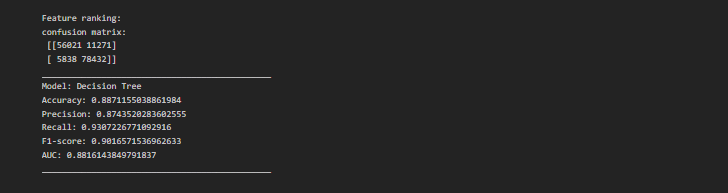


## Decision tree

Our feature reduction model was implemented to detect the variables which most impacted customer churn which was then applied to a decision tree classification model. After ranking variables based on importance scores, variables that are important are mostly the same as those before. Therefore, they are selected as predictors.

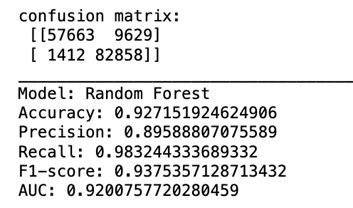


The metrics that were yielded by the model indicated that it was reliable and accurate. Our true negatives correctly identified 56,021 customers as churning, 11,271 customers incorrectly predicted to churn (false positives), 5,838 customers who churned but were not predicted to do so (false negatives) and 78,432 customers were correctly identified as churning (true positives). Our decision tree also yielded the following performance metrics: a 88.71% accuracy which indicates the model is effective in identifying churn in the majority of cases. Our precision was scored at 87.44% which suggests that when the model predicts a customer is churning, it identifies 87.44% of those customers who might be a target client base. Recall came in at 93.07%. The model’s F1 score came in at 90.17% which suggests the comprehensiveness of the model and finally, it had an AUC score of 88.16% which indicates excellent model performance. The decision tree model demonstrated an excellent capability at looking at different features and identifying factors that will lead to customer churn and identifying them before the customer churns. This will allow us to implement recommendations before the custom churns to keep them.

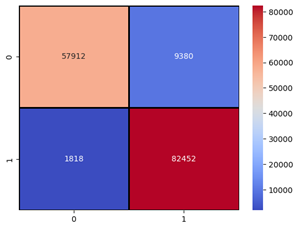
****

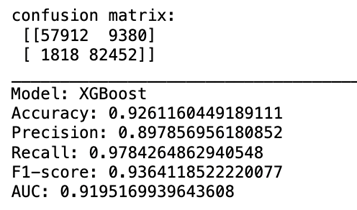
We also tried displaying the decision tree using Python and Pydot as below.

****

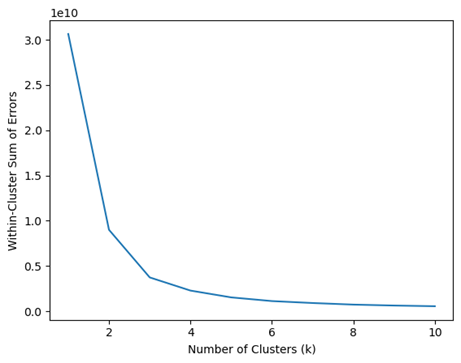
Next, a random forest model with n equal to 50 is built. The model provides both accuracy and F1-score of around 0.93.

## XGBoost

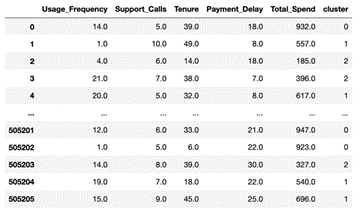
We also tried running the XGBoost model with results of the confusion matrix as follows.



## Clustering

****

Based on Usage\_Frequency, Support\_Calls, Tenure, Payment\_Delay, and Total\_Spend, we clustered customers after estimating the most compatible number of clusters, which was 3.

A graph of a heat map

Description automatically generated with medium confidence

# Recommendations:

In addressing the challenges of customer churn, our initial recommendation centers around enhancing engagement with customers facing multiple support calls and technical challenges. To effectively reduce the risk of churn, the step involves a deep analysis of customer support data, looking into the specifics of support calls to identify pain points and recurrent technical issues. The systematic categorization of support calls, based on, frequency, and resolution time, becomes important in understanding patterns contributing to customer dissatisfaction. This sets the stage for personalized support, where assistance is tailored to address the unique concerns of customers grappling with more difficult technical issues. Implementing a tiered support system further amplifies this approach by prioritizing customers with more complex issues and ensuring special attention. By acknowledging and addressing the specific needs of customers navigating technical issues, businesses demonstrate a commitment to individualized care, markedly increasing the likelihood of prompt issue resolution, fostering customer loyalty, and curbing the threat of churn.

## Personalized Support

The first recommendation we suggest is to enhance engagement with customers experiencing elevated support calls and technical challenges. In order to address customer churn, the first step involves a detailed analysis of customer support data. This entails exploring details of support calls to discern specific pain points and frequent technical issues encountered by customers. By analyzing previous support interactions, businesses can uncover patterns, trends, and recurring problems that contribute to customer dissatisfaction. Categorizing support calls becomes imperative, involving the systematic classification of issues based on their nature, frequency, and the time taken for resolution. Tailoring assistance to address the unique concerns of customers facing higher technical complexities establishes a more empathetic and effective support system. A system that can help with amplified customer service is implementing a tiered support system, where customers with more complex issues receive priority attention. This tiered structure acknowledges the varying degrees of difficulty in customer concerns and ensures that those facing the most challenging problems benefit from specialized assistance. By acknowledging and catering to the specific needs of customers facing technical challenges, businesses demonstrate a commitment to individualized care, significantly increasing the likelihood of resolving issues promptly and satisfactorily, thereby fostering customer loyalty and mitigating the risk of churn.

## Increase in Customer Retention

Enhanced Customer Loyalty involves the implementation of loyalty programs that extend beyond transactional exchanges. These rewards offer customers tangible rewards, such as exclusive discounts, free products, or personalized experiences, expressing a genuine appreciation for their continued patronage. The analysis of customer churn data unveils comprehensive customer profiles, enabling businesses to discern trends, preferences, and evolving needs within their customer base. By gaining insights, companies can tailor offerings and communication strategies, injecting a personalized touch into loyalty programs.

Providing rewards tailored to each customer allows companies to build a personal connection between the customer and the company, based on individual preferences fostering a sense of recognition. This allows for customers to provide positive feedback, and reinforces customer loyalty as ongoing benefits.

# Faith and Ethics Implications

When analyzing customer churn data, ethical practices play a crucial role in shaping the methodologies used throughout the process and outcomes derived. Especially, the ethical implications surrounding privacy and data security, bias and fairness, while having a unique lens on Christian ethics. Striving to protect customer information, companies must adhere to regulations and laws, seeking explicit consent for data usage while providing opt-out options. By integrating certain practices of trustworthiness, fairness, and stewardship, we underscore the ethical foundations that should be considered when dealing with customer data.

## Privacy and Data Security:

When dealing with consumer data, it is essential to protect and maintain confidentiality with customer data. Protecting data privacy involves adhering to strict regulations and laws depending on the organization, to safeguard customer information throughout the analysis process. When obtaining explicit consent for data usage, we have to clearly communicate the purposes of data handling, and allow customers the option to opt out. This ensures the protection of customer privacy. Proper data security measures must be implemented to protect the dataset against unauthorized access, this ensures that sensitive customer details remain confidential. Such security measures include encryption protocols, access controls, and secure storage practices to reduce the risk of data breaches. By prioritizing both data privacy and security, companies can build and maintain trust with their customers while adhering to legal and ethical standards in handling sensitive information.

## Bias and Fairness

Addressing bias and promoting fairness is key in customer churn analysis. Algorithmic bias poses a significant ethical concern, since it can arise from historical imbalances in the data used for training models. It is essential to be careful and aware of potential biases that may exist in the dataset or be introduced during the analysis. For example, if the training data reflects historical imbalances, such as preferential treatment for certain subscription types, the resulting model may perpetuate these biases, impacting the fairness of its predictions. By acknowledging the existence of such biases, we are able to take the first step in addressing them, allowing us to refine the analytical process to ensure fair outcomes. Additionally, identifying biases is imperative to actively work towards fair treatment throughout the analysis. This involves adjusting the decision-making processes to reduce any unwarranted advantages or disadvantages experienced by specific demographics or groups of customers. By prioritizing fair treatment, businesses can enhance the ethical standing of their analyses and also foster inclusivity and trust among their diverse customer bases.

## Christian Ethics

This project focuses on diminishing customer churn which aligns with Christian teachings on responsible stewardship - 1 Corinthians 4:2, "Now it is required that those who have been given a trust must prove faithful." In this context, the data represents the entrusted information about individuals. This verse emphasizes the importance of trustworthiness in stewardship, and this theory should serve as a guide for ethical considerations associated with leveraging insights from the data. Christian ethics highlights the responsibility of companies as stewards, this emphasizes companies to handle customer information with integrity and transparency, ensuring that the strategies derived from the analysis are trustworthy. Additionally, the concepts laid out in 1 Corinthians 4:2 imply that stewards have to be faithful and true to their actions. Therefore, the strategies implemented to retain customers should not exploit individuals or contribute to social inequalities. By emphasizing the importance of trustworthiness, fairness, and compassionate stewardship, the customer churn analysis can align more closely with Christian ethical principles, creating a foundation for sustainable business practices that resonate with the teachings of responsible stewardship.