Customer Churn Prediction

A Data-Driven Analysis for Subscription-Based Businesses

# Table of Contents

[Table of Contents 1](#_wjgge4wx0h7t)

[Abstract 2](#_zfsjrc28zg95)

[**1. Introduction 2**](#_n079wkr37snx)

[1.1. Background 2](#_5nltayz88syb)

[1.2. Motivation 2](#_qyxyaqv9kde3)

[1.3. Project Objectives 2](#_4rcwkve6lwfo)

[**2. Methodology 3**](#_xp43fjxfv7ps)

[2.1. Data Source and Description 3](#_95gddrtf9za2)

[2.2. Data Preprocessing 3](#_ia3qdgim3jco)

[2.2.1. Data Combining 3](#_gugr4x71ovn4)

[2.2.2. Data Transformation 3](#_dcsppr5lt4wo)

[2.2.2.1. Categorical Encoding: 3](#_oe9t5jsuyoz6)

[2.2.2.2. Data Correction: 3](#_q24koln38gc)

[2.3. Feature Engineering 3](#_mnm503glp52v)

[2.4. Model Selection 4](#_b93irphcu0s2)

[2.5. Model Training and Evaluation 4](#_ycfm3dojssio)

[2.5.1. Data Splitting 4](#_falsxex6xvro)

[2.5.2. Model Evaluation and Hyperparameter Tuning 4](#_eau67779g79q)

[2.6. Feature Importance Analysis 4](#_ucoi0ysmijk2)

[2.7. Interpretation Techniques 4](#_e1zonxfqr2nm)

[2.7.1. Partial Dependence Plots (PDP) 4](#_tbr7nu2rpl5b)

[2.7.2. SHAP (SHapely Additive exPlanations) 5](#_cpu0hfbh941q)

[2.8. Initial Data Analysis 5](#_76rxhztvv82a)

[2.8.1. Churn Distribution 5](#_1kmn6i9d8n5f)

[[Fig. 1 - Distribution of Churn] 5](#_iw8rcr8pbhle)

[2.8.2. Income Distribution by Churn 6](#_14xhinnrvx0h)

[[Fig. 2 - Income Distribution of Churn] 6](#_ih4mfpjsma0q)

[2.8.3 CLV by Customer Segment 7](#_qhurw6hp58dy)

[[Fig. 3 - Customer Lifetime Value by Customer Segment (Income)] 7](#_mepcix14l7cw)

[[Fig. 4 - Customer Lifetime Value by Customer Segment (Income)] 7](#_m0wi7pbzcrju)

[**3. Results and Discussion 8**](#_ct4hocp8vi3t)

[3.1. Model Performance 8](#_9ie6x1kim6hf)

[[Fig. 1 - Confusion Matrix] 8](#_pbbpo8apoxo)

[3.2. Key Drivers of Churn 9](#_4kh9odl8ptoc)

[[Fig. 2 - Feature Importances] 9](#_un2l6njfs7dn)

[3.3. Insights from Partial Dependence Plots: 10](#_yztcd539em0b)

[[Fig. 3 - Partial Dependence Plots] 10](#_tsq328jg822h)

[3.4. SHAP Analysis 11](#_a5q826nmrh2g)

[[Fig. 4 - SHAP Summary Plot] 11](#_k5mnmnccgoj8)

[[Fig. 5 - SHAP Force Plot] 11](#_cuhap2vgly8s)

[**4. Implications and Recommendations 12**](#_r6238g34ntoo)

[4.1. Actionable Areas 12](#_avg8x5w1ssp7)

[4.2. Socioeconomic Effects of Churn 12](#_u8ypvuam02jk)

[**5. Limitations and Future Work 12**](#_yifiuc53ryca)

# Abstract

Customer churn, defined as losing subscribers over time, poses a significant challenge for subscription-based businesses, directly impacting revenue and profitability. This paper explores the development and analysis of a machine-learning model to predict customer churn for Microsoft’s Contoso. Utilizing provided customer data, we employ random forest classification, a well-suited method for handling non-linearity and addressing overfitting concerns. In-depth analysis using Partial Dependence Plots (PDP) and SHAP (Shapley Additive exPlanations) techniques offer valuable insights into the key drivers influencing churn, enabling the formulation of targeted customer retention strategies.

# 1. Introduction

## 1.1. Background

Subscription-based businesses thrive on customer retention, and involuntary customer churn significantly hinders their growth and financial stability. Churn refers to the phenomenon where subscribers terminate their subscriptions, leading to revenue loss and decreased customer lifetime value.

## 1.2. Motivation

Predicting customer churn proactively empowers businesses to develop targeted customer retention strategies. By identifying individuals at an elevated risk of churning, companies can implement interventions to mitigate churn and maintain a healthy subscriber base. Machine learning (ML) offers a powerful approach to churn prediction, as it can discover complex patterns and relationships within customer data that traditional methods might miss.

## 1.3. Project Objectives

This project aims to:

1. Develop a Machine Learning Model for Churn Prediction: We will construct a random forest classification model to predict customer churn based on historical data.
2. Identify Key Drivers of Churn: Employ feature importance analysis and interpretation techniques to pinpoint the factors most significantly influencing customer churn.
3. Derive Actionable Insights for Churn Mitigation: Leverage the model's insights to formulate customer retention strategies that target high-risk segments and address the underlying factors contributing to churn.

# 2. Methodology

## 2.1. Data Source and Description

The dataset for this project incorporates customer data from three sources:

* Contact.csv: This file contains customer demographics and contact information.
* Customer-Loyalty.csv: Includes details about customer loyalty program membership, reward points, and loyalty tiers.
* SubscriptionHistoryContoso.csv: Encompasses customer subscription details, including subscription amount, duration, type, and start/end dates.

## 2.2. Data Preprocessing

### 2.2.1. Data Combining

The data from the three sources was merged using inner joins on the CustomerID field to create a comprehensive customer profile dataset.

### 2.2.2. Data Transformation

#### 2.2.2.1. Categorical Encoding:

Categorical features were encoded using one-hot encoding to transform them into numerical representations suitable for the machine learning model.

#### 2.2.2.2. Data Correction:

Several columns that would hurt the prediction accuracy of the churn model were removed.

## 2.3. Feature Engineering

New features were derived from existing data to enhance model performance and gain deeper insights potentially:

* TotalSubscriptionLength (calculated by subtracting the subscription start date from the end date)
* CustomerAge (calculated based on the birthdate provided in Contact.csv)
* LoyaltyAge (calculated based on the loyalty start date supplied in Customer-Loyalty.csv)
* AccountAge (calculated based on the account start date provided in Customer-Loyalty.csv)
* NumberOfSubscriptions (total number of subscriptions a customer has)

## 2.4. Model Selection

A random forest classifier was chosen for churn prediction due to its:

* Effective handling of non-linear relationships between features and the target variable (churn).
* Robustness to overfitting, which is crucial to avoid the model performing well on the training data but poorly on unseen data.

## 2.5. Model Training and Evaluation

### 2.5.1. Data Splitting

The data was split into training and testing sets using a stratified sampling approach to maintain class balance (i.e., ensuring approximately the same proportion of churned and non-churned customers in both sets). A standard split ratio of 80% for training and 20% for testing was employed.

### 2.5.2. Model Evaluation and Hyperparameter Tuning

Several evaluation metrics were used to assess model performance and guide hyperparameter tuning:

* Accuracy: The proportion of correctly classified cases (churn and non-churn).
* Precision: Measures what proportion of the customers predicted as churn actually churned.
* Recall: The proportion of actual churn customers the model correctly identified.
* F1-score: A balanced metric considering both precision and recall.
* ROC-AUC score: Illustrates the model's ability to discriminate between churn and non-churn customers across all classification thresholds.
* Confusion Matrix: A breakdown of true positives, false positives, true negatives, and false negatives for a comprehensive view of the model's performance.

## 2.6. Feature Importance Analysis

The random forest model's built-in feature importance attribute was used to determine the features contributing most to its predictive power. The feature importance scores represent how much each feature contributes to decreasing the model's impurity (Gini impurity in this case).

## 2.7. Interpretation Techniques

### 2.7.1. Partial Dependence Plots (PDP)

PDPs illustrate the average marginal effect of individual features on the model's predictions. PDPs were generated for the top features identified as the most important to visualize their relationship with churn probability.

### 2.7.2. SHAP (SHapely Additive exPlanations)

SHAP analysis was applied for model interpretability on local (individual customer) and global (feature importance over the entire dataset) levels. SHAP values attribute predictions to each feature, helping to understand the model's rationale.

## 2.8. Initial Data Analysis

### 2.8.1. Churn Distribution

### 

##### [Fig. 1 - Distribution of Churn]

This figure is crucial to the analysis of this dataset. From this figure, we can see that the number of customers who did not churn is substantially higher than those who did. This means that the business model is already in a decent position for growth, but modeling churn may still give insight into how to lower the number of ‘Yes’ further.

### 2.8.2. Income Distribution by Churn

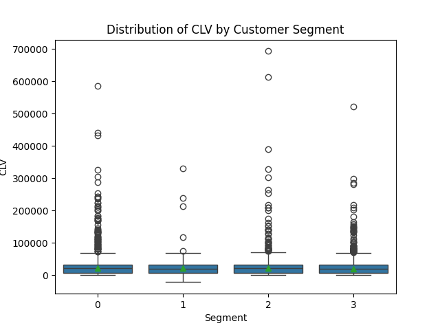
### 

##### [Fig. 2 - Income Distribution of Churn]

This figure shows the distribution of customers who churned with respect to income. This plot shows that lower-income individuals, specifically those making between $25,000 and $37,500 per year, are at a much higher risk of churn, with approximately 12,000 subscriptions churning in that bracket.

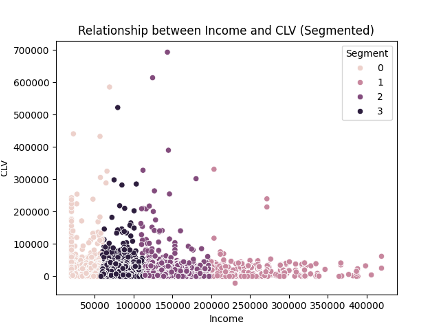
### 

### 2.8.3 CLV by Customer Segment



##### [Fig. 3 - Customer Lifetime Value by Customer Segment (Income)]

Figure 3 shows four box plots that break up the customer lifetime value into four income segments. What can be seen here is that customers in the 50th to 75th percentile segment of customer income have the highest CLV ROI values of the four segments. Interestingly, the highest segment does not have the highest CLV, and in terms of the density of spread, the lowest income segment has the highest consistent CLV.



##### [Fig. 4 - Customer Lifetime Value by Customer Segment (Income)]

Figure 4 supports this claim by showing the consistently low CLV of the high-income segment. While they may be capable of affording more expensive subscriptions, they do not provide a high value to the company. On the other hand, customers making $150,000 or less per year are the most dense and robust in terms of the data, with the highest value being seen in customers below $100,000 in segments 1 and 2.

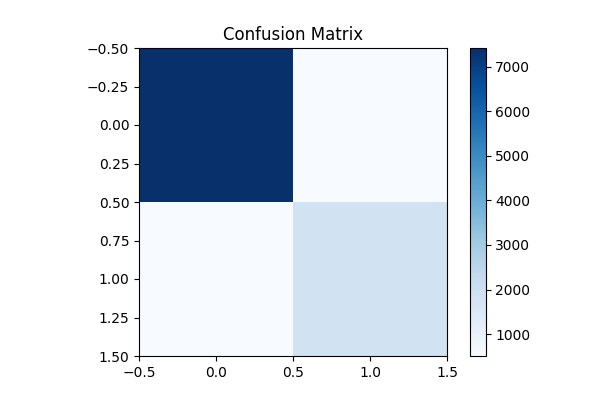
# 3. Results and Discussion

## 3.1. Model Performance

The final random forest model, after hyperparameter tuning, achieved the following performance metrics on the held-out testing set:

* Accuracy: 89.89%
* Precision: 77.91%
* Recall: 77.64%
* F1-score: 77.77%
* ROC-AUC score: 0.86

The confusion matrix (Fig.1) provides further insights into the model's accuracy in classifying both churned and non-churned customers.

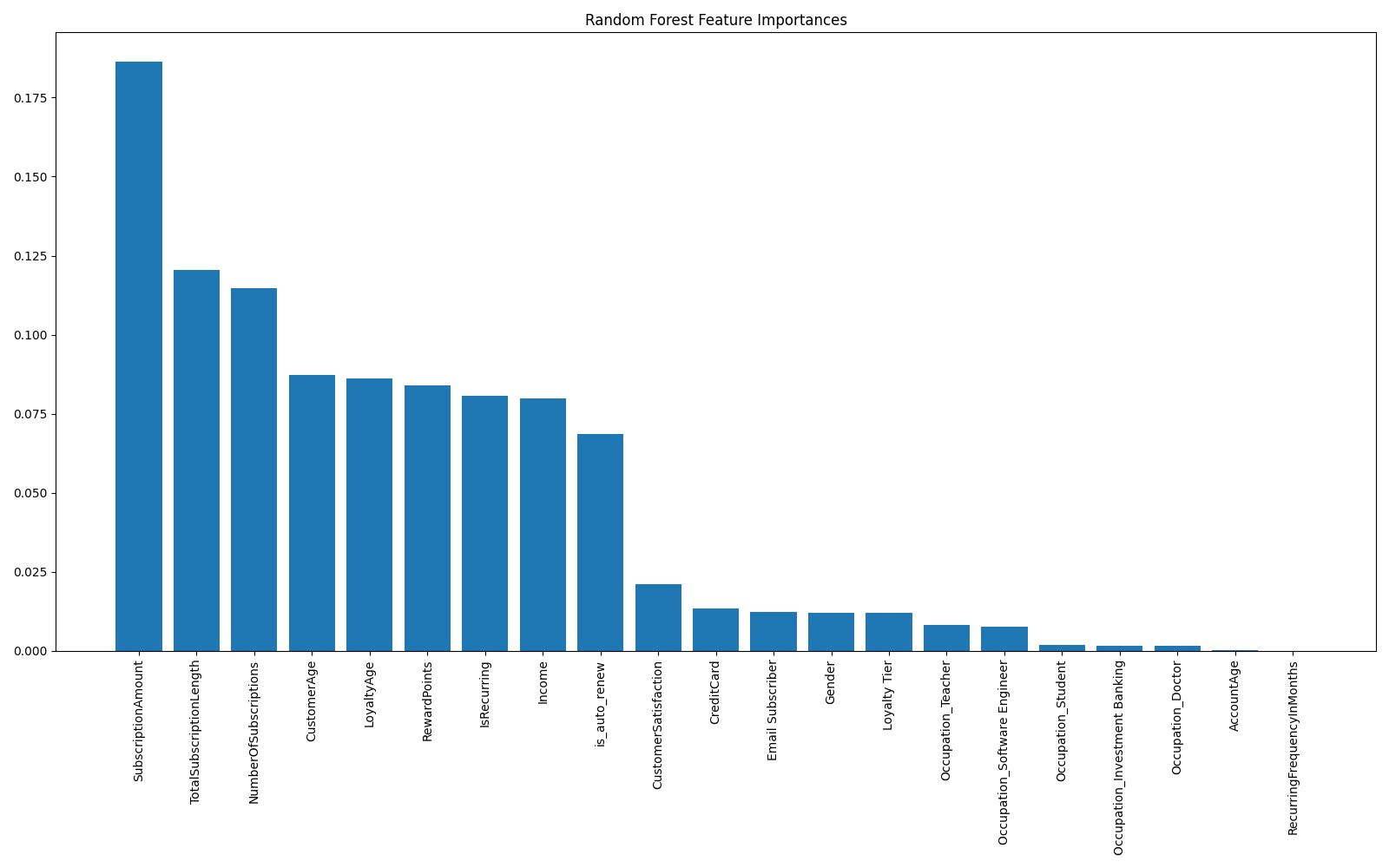


##### [Fig. 1 - Confusion Matrix]

## 

## 3.2. Key Drivers of Churn

Feature importance analysis (Fig. 2) revealed the following features as the top predictors of churn:



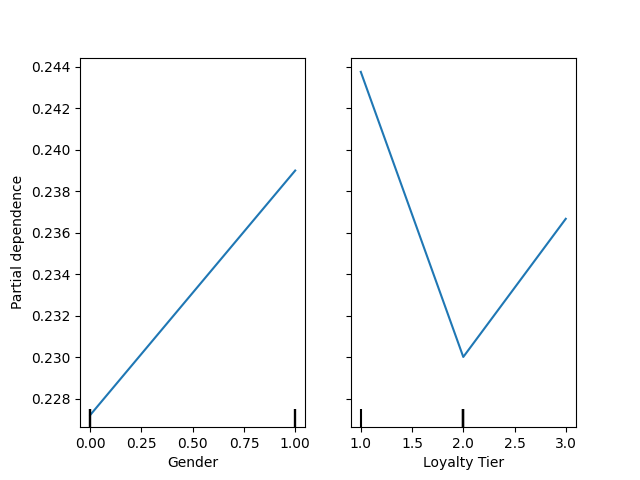
##### [Fig. 2 - Feature Importances]

* Subscription Amount: Customers paying lower amounts exhibit a higher churn risk.
* Total Subscription Length: Longer subscriptions are associated with decreased churn probability.
* Number of Subscriptions: A higher number of subscriptions implies a lower churn likelihood.
* Customer Age: Older customers appear slightly less prone to churn.
* Loyalty Age: Shorter loyalty program membership durations show a higher correlation with churn.

## 

## 3.3. Insights from Partial Dependence Plots:

* Gender: The PDP (Fig. 3) indicates a marginal difference in churn probability between genders, with male customers potentially having a slightly higher churn risk.
* Loyalty Tier: The PDP (Fig. 3) indicates a significant difference in the churn risk of low-tier loyalty members versus high-tier customers. It also suggests that tier 2 has the lowest churn rate, with tier 1 having the highest churn.



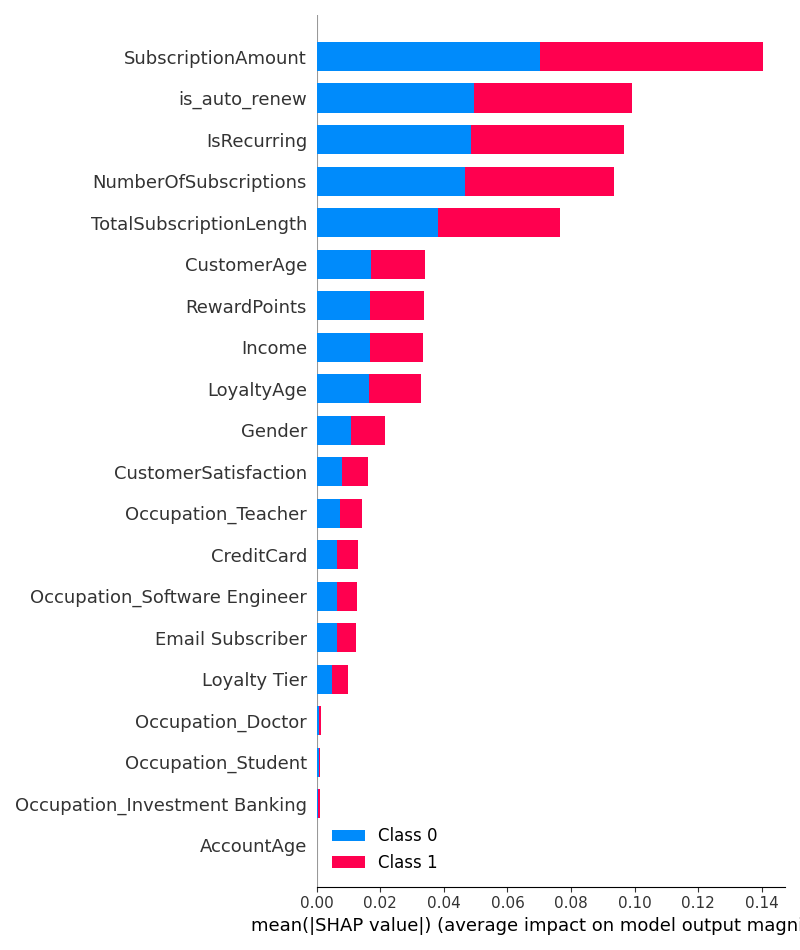
##### [Fig. 3 - Partial Dependence Plots]

Caveats: It is crucial to avoid overgeneralizing based on these plots and to remember that correlation does not necessarily imply causation.

## 

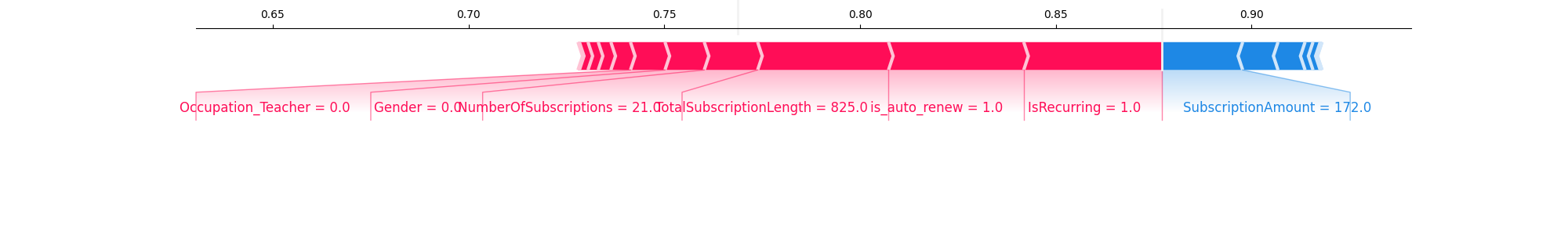
## 3.4. SHAP Analysis

The SHAP summary plot (Fig. 4) highlights the top influential features and their directional impact on churn prediction. It shows that the subscription price has the highest effect on customer churn rate. Combined with information from other analyses and charts, it also tells that customers who have auto-renew enabled are less likely to churn. It shows that the number of historical subscriptions and the total subscription history with the business significantly affect churn.



##### [Fig. 4 - SHAP Summary Plot]

Individual SHAP force plots (e.g., Fig. 5) were used to understand specific customer-level predictions in a personalized manner. They showed how different features affected the churn rate. This plot reinforces the abovementioned point regarding subscription price and its effect on customer churn rate. It also reinforces that an auto-recurring subscription makes a customer much less likely to churn.



##### [Fig. 5 - SHAP Force Plot]

# 4. Implications and Recommendations

## 4.1. Actionable Areas

The findings highlight actionable areas for subscription-based businesses to focus on for churn reduction:

* Tiered Pricing: Introduce flexible subscription tiers with varying pricing to appeal to a broader customer base and reduce churn due to cost.
* Targeted Promotions: Offer renewal incentives or loyalty discounts to customers with high churn risk, particularly those with shorter subscription lengths and fewer subscriptions.
* Proactive Support: Prioritize customer support for those recently enrolled in loyalty programs to create positive early experiences and foster brand loyalty.
* Feature Enrichment: Explore additional features that might improve model performance, such as customer engagement metrics and support interaction data.
* Addressing Bias: Remain aware of potential bias in the data and the model. Investigate fairness metrics to avoid inadvertently discriminating based on factors such as gender.

## 4.2. Socioeconomic Effects of Churn

Calculating and reducing churn can have several benefits for a business. It can allow a company to increase the value per customer, reduce CAC, and increase overall shareholder value. However, one key aspect of churn that is often overlooked is how it affects the relationship between the business and the customer.

For a customer, reducing churn means spending hard-earned money on a service a business provides without fail each month. If churn is down, the relationship between the customer and the company is strong, and the business has built rapport with the customer. The analysis above shows a clear correlation between the subscription price and churn, as well as income and churn. Through this, we can see that those in lower tax brackets are significantly more likely to churn than those in higher tax brackets. This shows that decreasing the price is a clear path to reducing churn. While this does not necessarily directly increase demand, it does increase CLV. So, despite keeping a consistent CAC, thereby losing a small portion of revenue in the short term, the increase in CLV and churn reduction will increase the income in the long term and show the customer that you listen to your market.

Overall, it is clear that lowering the price will make the product more accessible to lower-income individuals while also increasing the business's long-term revenue streams. This is a mutually beneficial project that brings the company closer to a true equilibrium price and quantity on the supply and demand curve.

# 

# 5. Limitations and Future Work

* Data Limitations: The dataset might not fully capture customer behavioral aspects, limiting the model's potential explanatory power.
* Evolving Churn Dynamics: Customer churn patterns can change over time.
* Model Generalizability: Caution is required when applying these findings to different business models, as factors influencing churn might vary.

# 6. Citations

[1] “How can I optimize the hyperparameters of this classifier for model improvement?” prompt. *Gemini*, version 1.0, Google, 9 Mar. 2024, gemini.google.com.

[2] Biau, Gérard, and Bin Yu. “Analysis of a Random Forests Model.” *Journal of Machine Learning Research*, vol. 13, 12 Apr. 2012.

[3] Joytaylor. “Subscription Churn Prediction Sample Guide - Dynamics 365 Customer Insights.” *Dynamics 365 Customer Insights | Microsoft Learn*, learn.microsoft.com/en-us/dynamics365/customer-insights/data/sample-guide-predict-subscription-churn. Accessed 9 Mar. 2024.

[4] Naik, Soham. “Churn Prediction- Commercial Use of Data Science.” *Analytics Vidhya*, 25 July 2022, [www.analyticsvidhya.com/blog/2021/08/churn-prediction-commercial-use-of-data-science/#:~:text=%27%20A%20common%20example%20is%20people,their%20usage%20of%20the%20service](http://www.analyticsvidhya.com/blog/2021/08/churn-prediction-commercial-use-of-data-science/#:~:text=%27%20A%20common%20example%20is%20people,their%20usage%20of%20the%20service).

[5] Paddle. “4 Steps to Predict Churn & Reduce Customer Attrition.” *Paddle*, 1 Sept. 2023, [www.paddle.com/resources/churn-prediction](http://www.paddle.com/resources/churn-prediction).

[6] “Random Forest.” *Corporate Finance Institute*, 22 Nov. 2023, [corporatefinanceinstitute.com/resources/data-science/random-forest/](http://corporatefinanceinstitute.com/resources/data-science/random-forest/).

[7] Shafi, Adam. “Random Forest Classification with Scikit-Learn.” *DataCamp*, DataCamp, 24 Feb. 2023, [www.datacamp.com/tutorial/random-forests-classifier-python](http://www.datacamp.com/tutorial/random-forests-classifier-python).

[8] Wiryaseputra, Michael. “Bank Customer Churn Prediction Using Machine Learning.” *Analytics Vidhya*, 12 Oct. 2022, www.analyticsvidhya.com/blog/2022/09/bank-customer-churn-prediction-using-machine-learning/.