January 13, 2025

# Churn Prediction

## Machine Learning Implementation

Muhammad Dhany Latief

# Preface

- This prediction aims to reduce customer churn rate, using the data of 4930 customers who have unsubscribed from our services, whether they churn or not.

- Churn will be predicted by using the existence of customers' dependencies, type of contract, months of tenure, which services they have subscribed to, the usage of paperless billing, and the amount paid per month.

- By utilizing this prediction model, hopefully the company can make good and informed decisions to make customer retention and acquisition efforts.

# Summary

# Executive Summary

## The average churn rate of Telecommunication Industry is said to be 31% as of 2024 (customergauge.com, 2025)

- We would like to differentiate customers that would churn and those who would not churn, because the **cost of acquiring new customers** is **five to 25 times higher** than the cost of **retaining existing customers**
- Churn is often caused by factors like poor customer service, high prices, or better offers from competitors

## Our company have faced a churn rate of 26% in total

- While still **below average**, ideally we'd like to **reduce** churn rate **as much as possible**
- We would like a prediction that can predict customers that would churn as much accurate as possible, preferably a prediction model that has the **least probability** of **predicting** customers that **churn** as customers that **do not churn**

## Using this prediction model, we might be able to predict which customers would churn more accurately compared to using arbitrary rules

- The implemented model would be able to **predict 95%** of the **actual churners** correctly
- This model is predicted to be able to **reduce** our **customer retention cost** by **73%** and **increase** our **net revenue** by **53%**

## We need more targeted customer retention efforts towards customers who are 'at risk' of churning

- Offer additional **deals** such as monthly charges discounts & bundle deals to **customers** who have only **used the services for 2 years or less**
- Offer **incentives** for **long-term contracts** such as loyalty rewards to **reduce churn** associated with **month-to-month contracts**

https://customergauge.com/blog/average-churn-rate-by-industry

# Business Metrics

- False Negative: When **customers** are **predicted to not churn**, but in reality, **they do churn**.
Consequence(s):

- **Revenue loss** from churned customers
- **Additional** expenditure in **acquisition cost** to replace churned customers

- False Positive: When **customers** are **predicted to churn**, but in reality, **they do not churn**.
Consequence(s):

- **Unnecessary** retention **costs** and efforts
- **Worsening customer experience** caused by **mistargeted retention efforts**

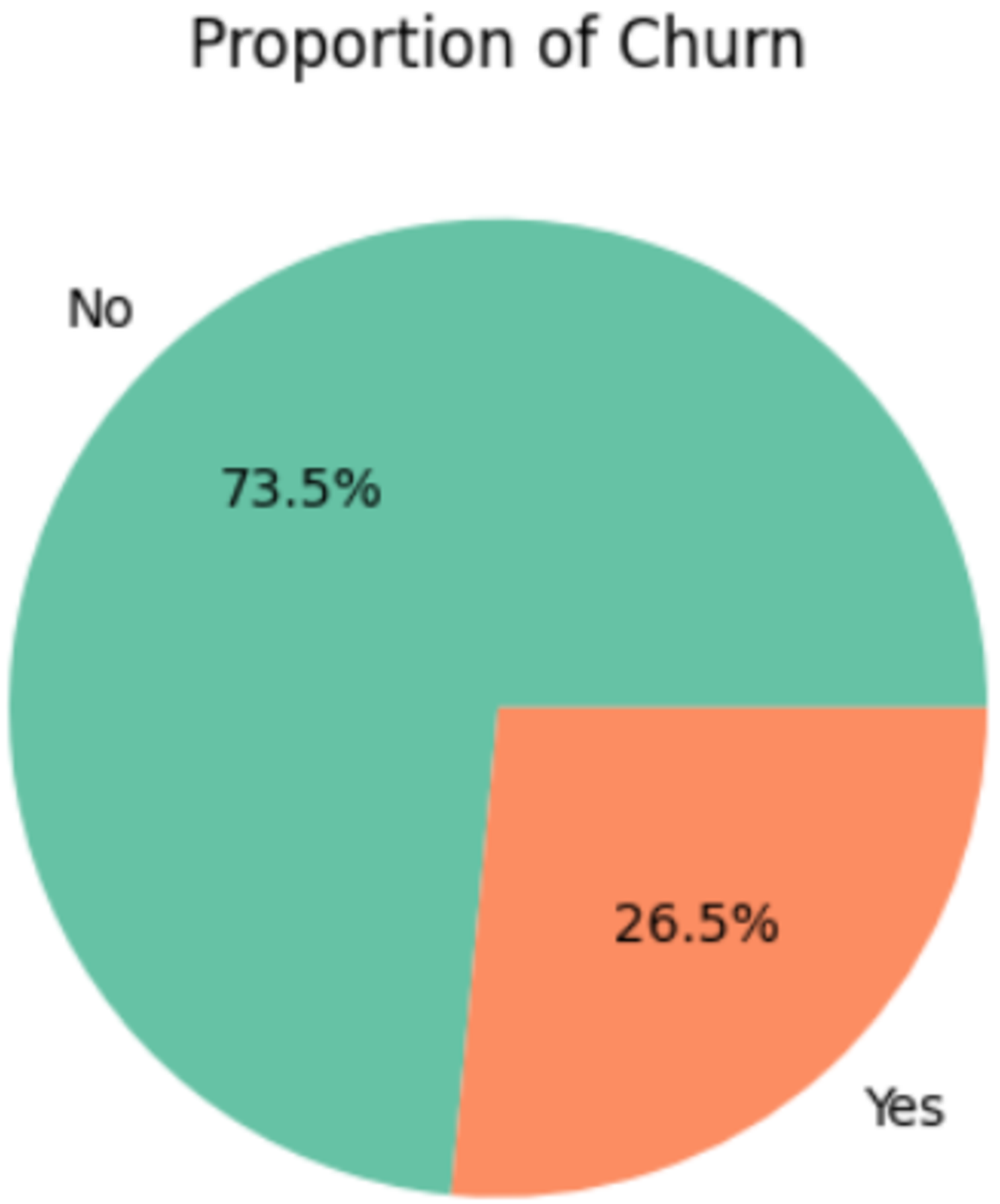
## As of 2024…

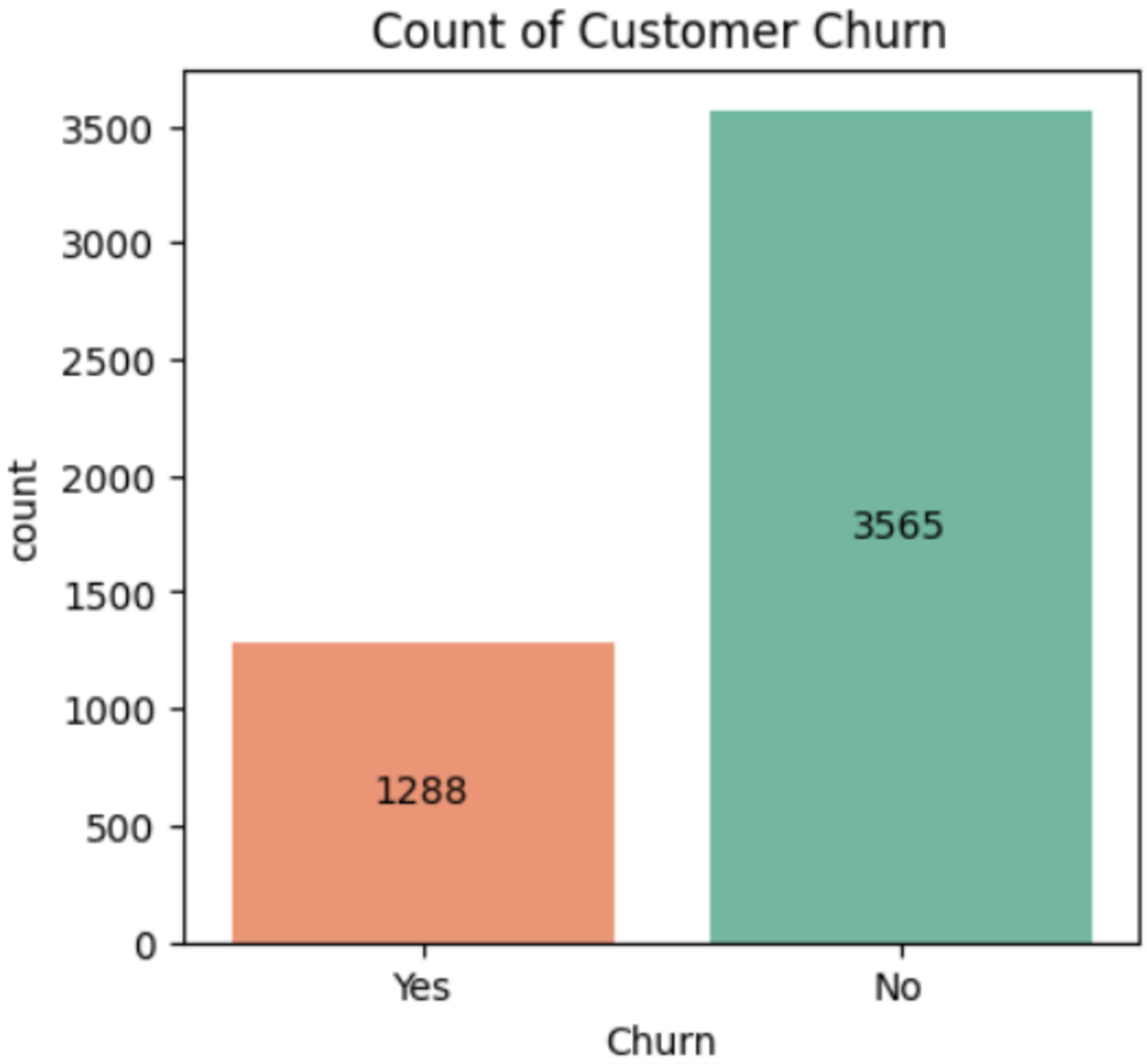
| Customer Acquisition Cost (CAC) | Customer Retention Cost (CRC) |
|---|---|
| **$620** /customer | **$125** /customer |

Based on these consequences, **False Negatives** will be **more costly** to the business as not only is there revenue loss, but the business also incurs high acquisition costs to replace churned customers.

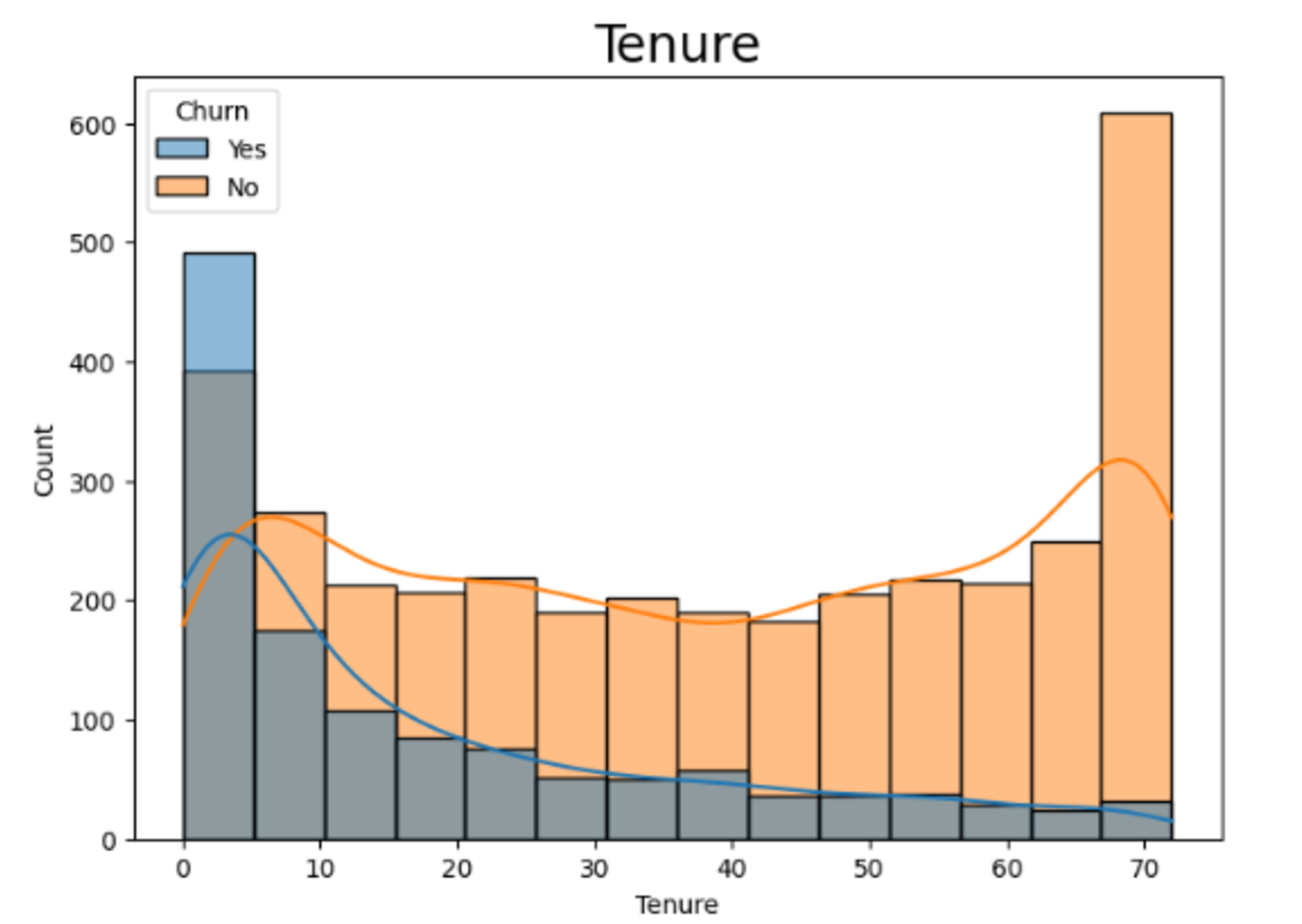
# Data Analysis

# Among 4930 customers who have left us, 1288 of them churned

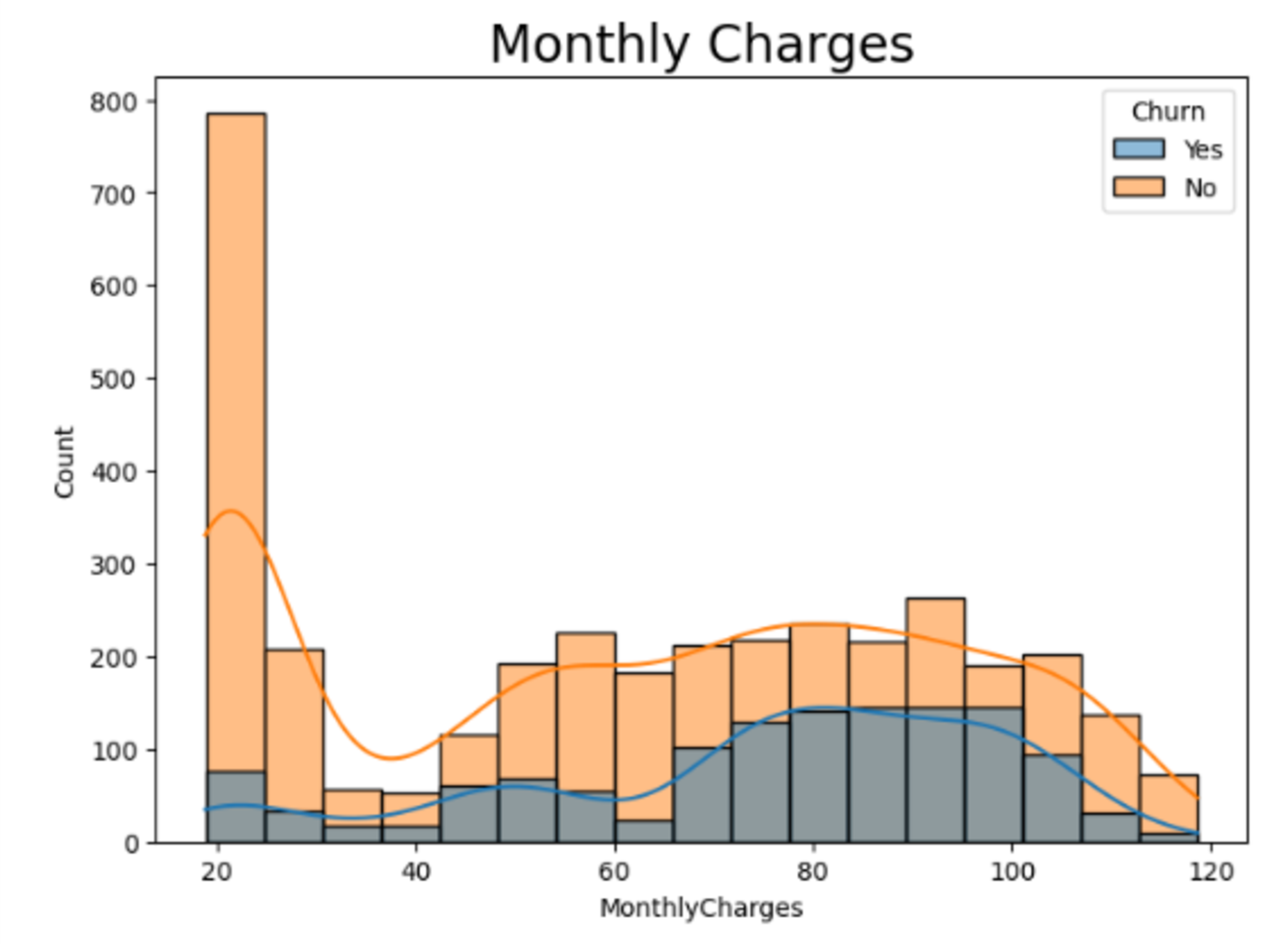


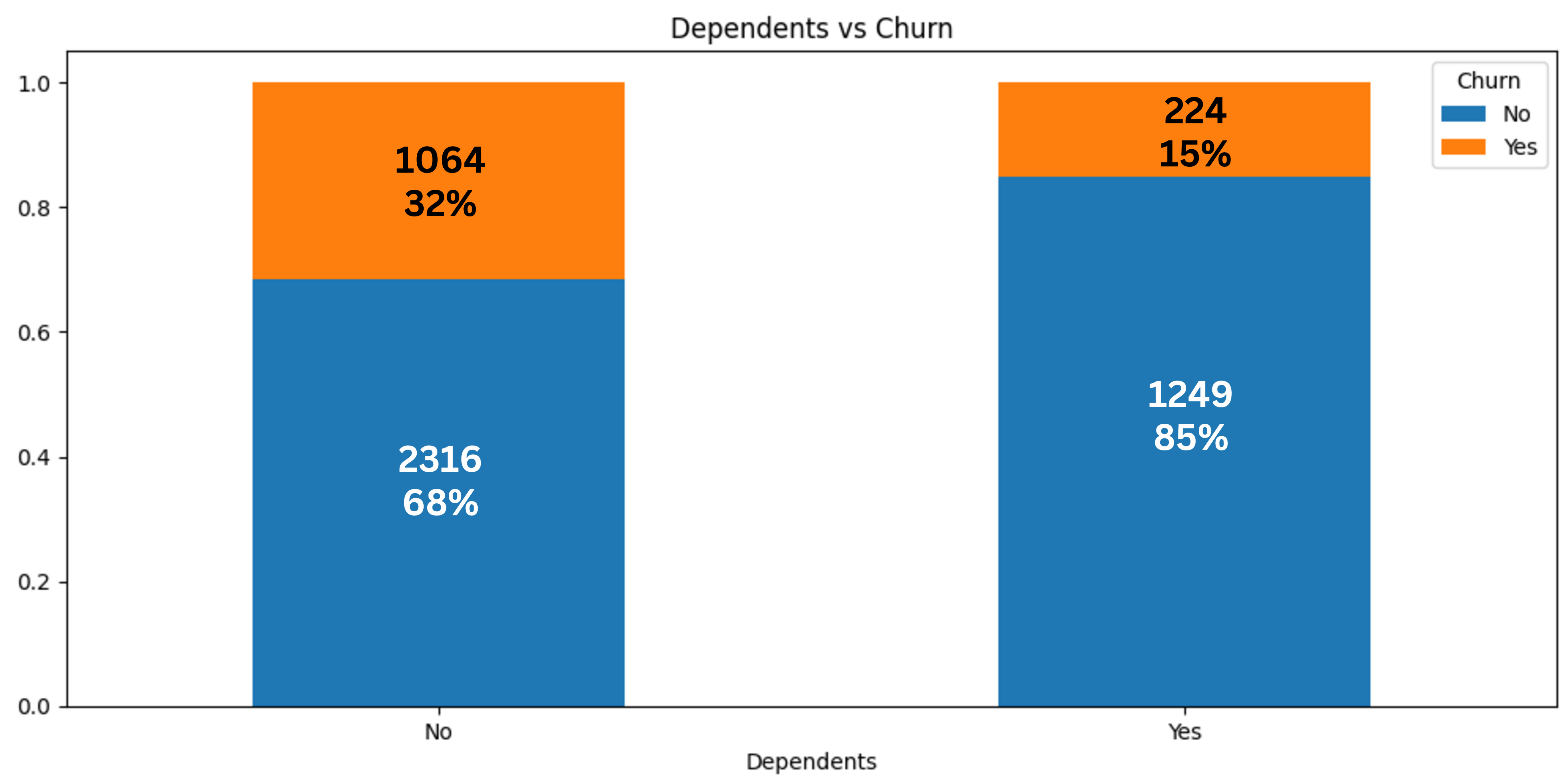
# Most churning customers have tenures of 0 to 5 months...



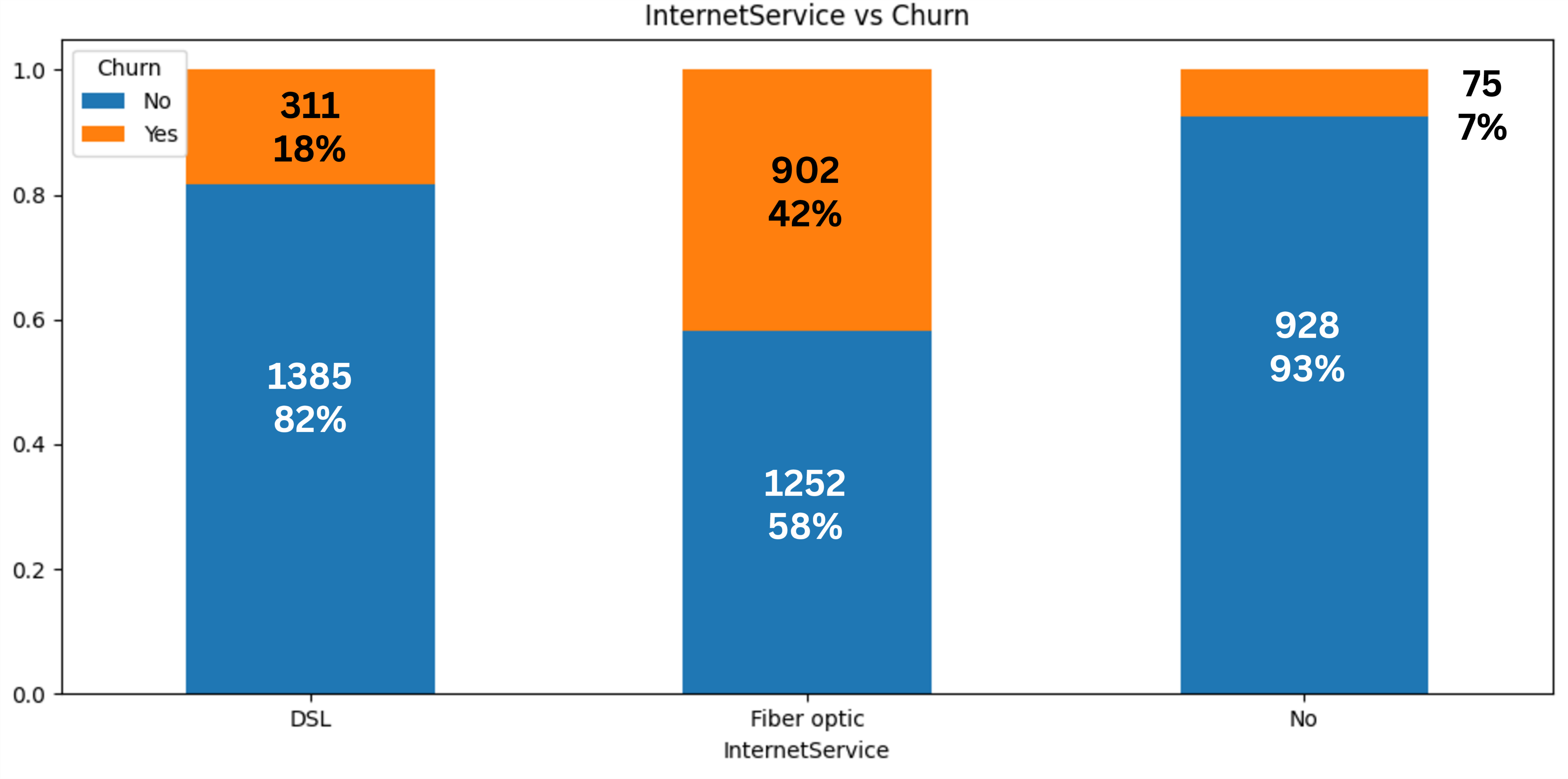
# ...and charged around $70 to $100 per month



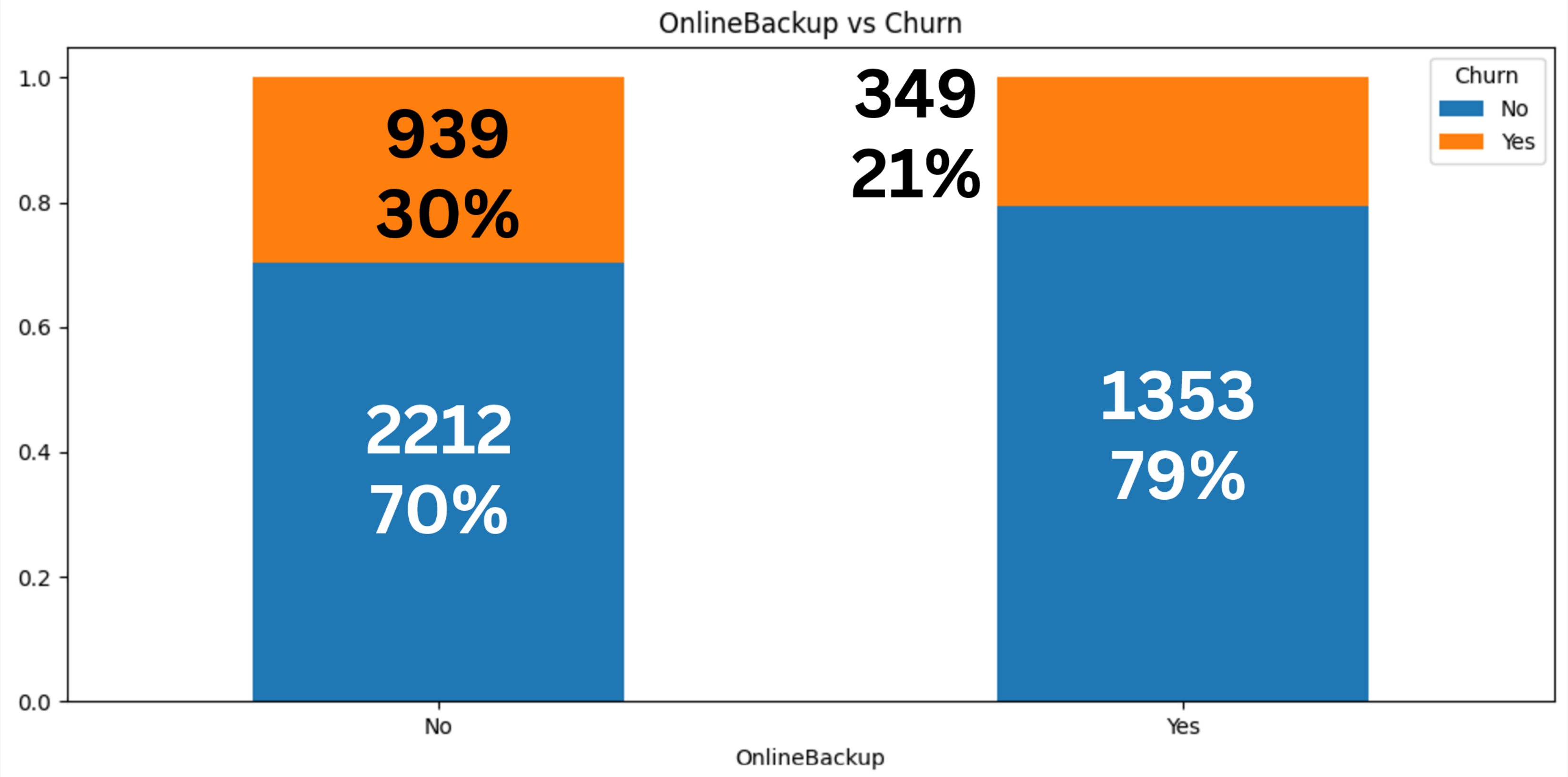
# Most customers who churned do not have registered dependents

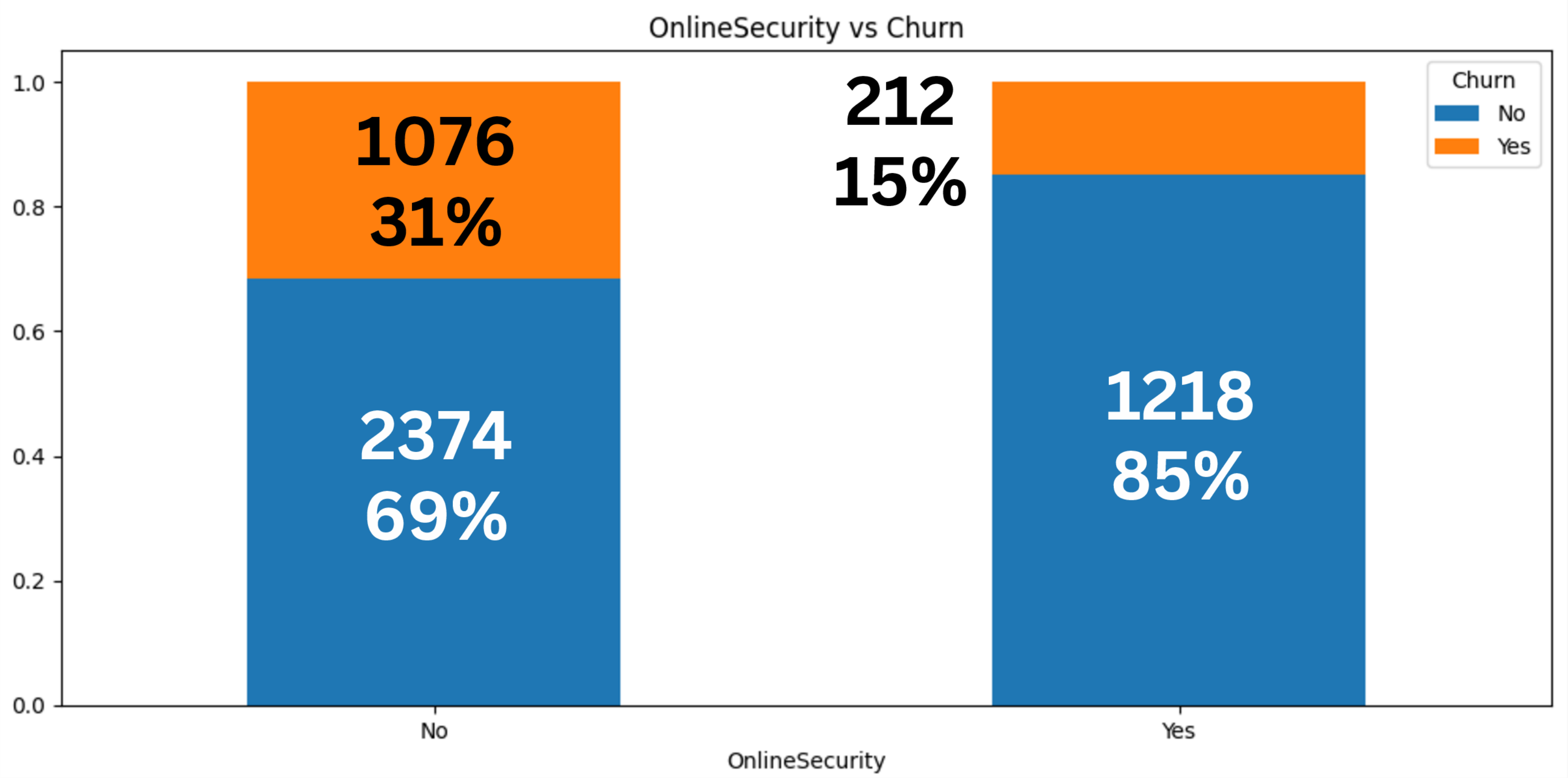


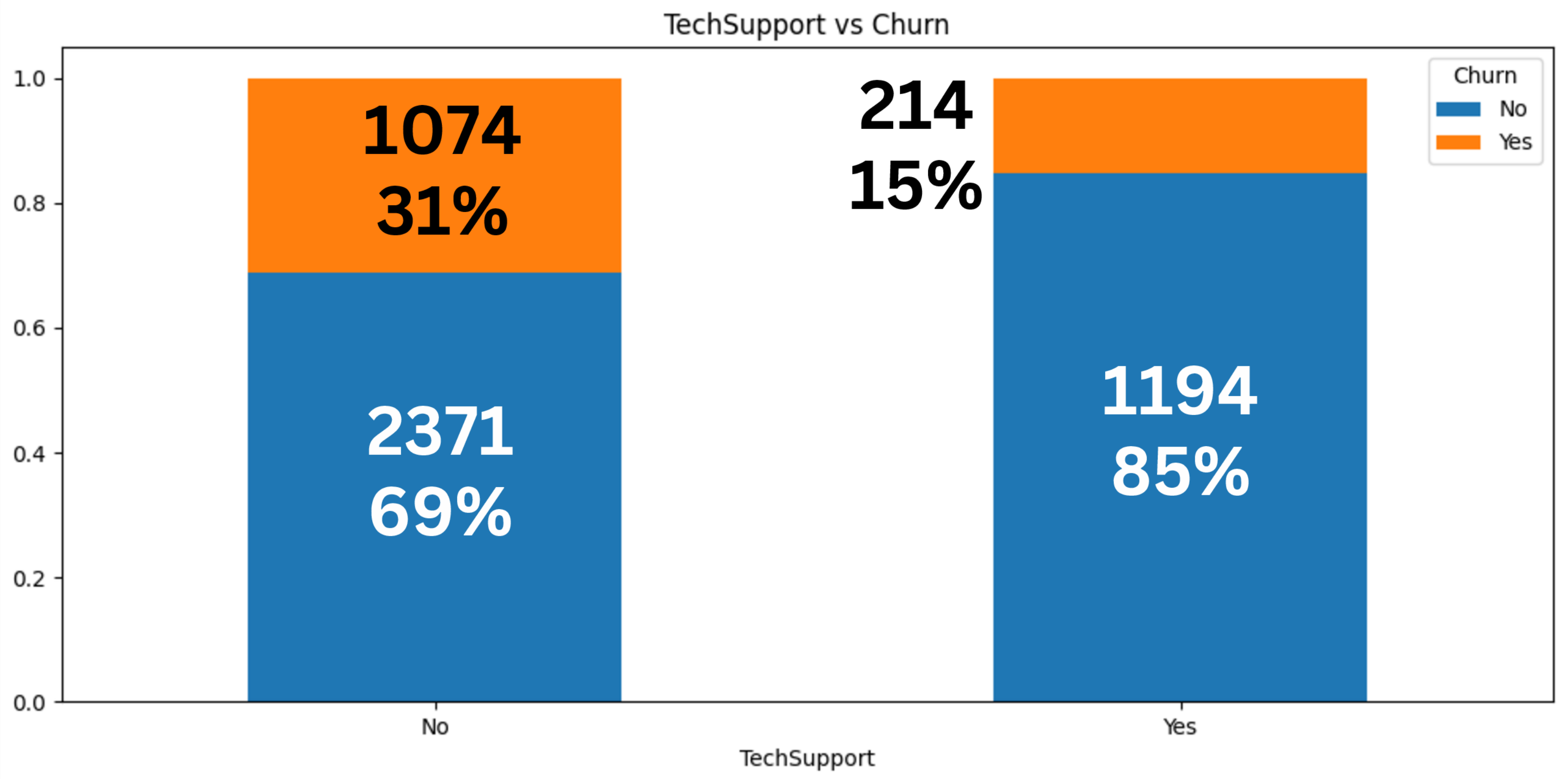
# Our fiber optic-based internet service has the highest churn rate

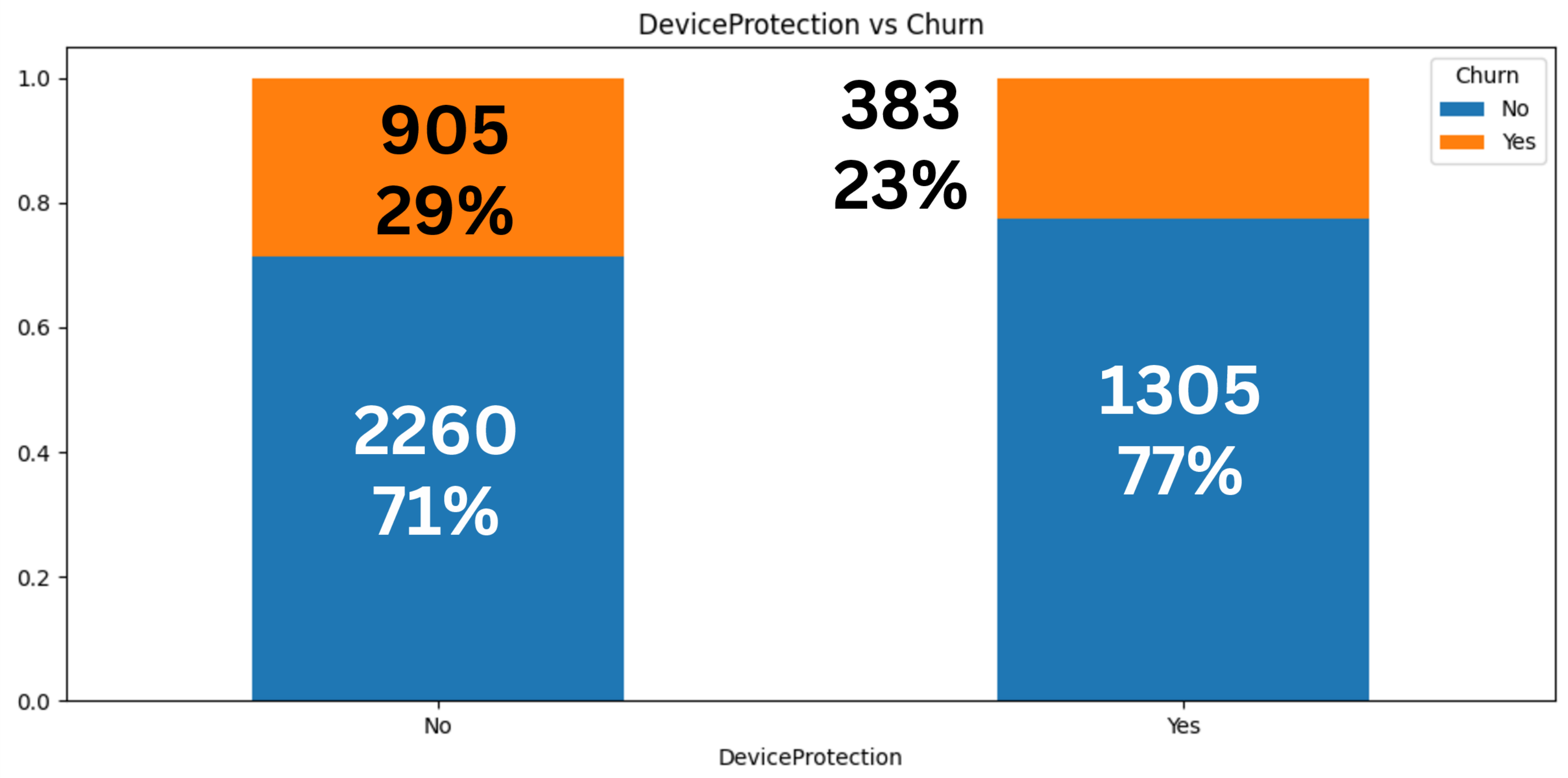


# Customers who do not subscribe to additional services related to the internet services tend to churn

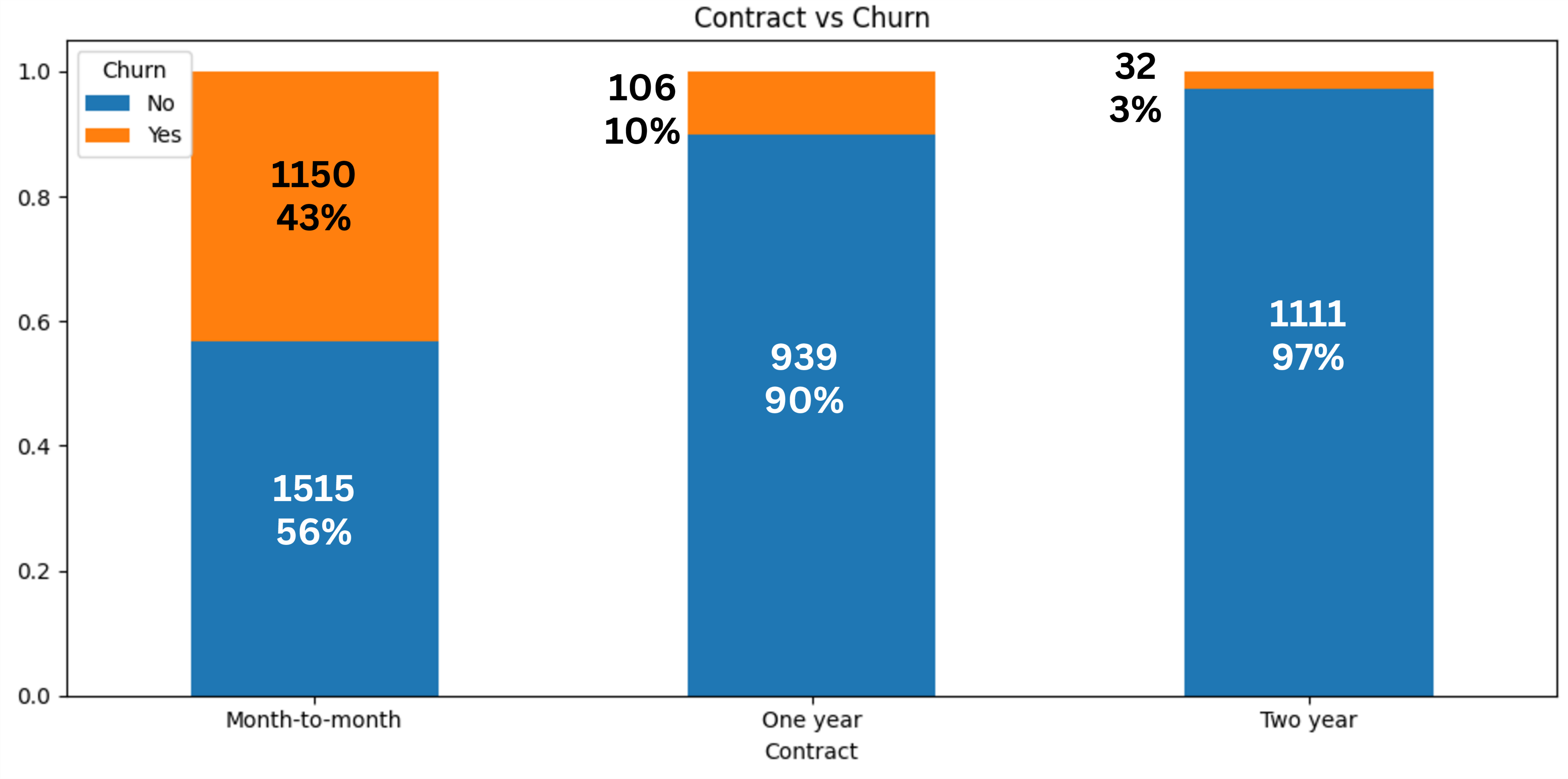


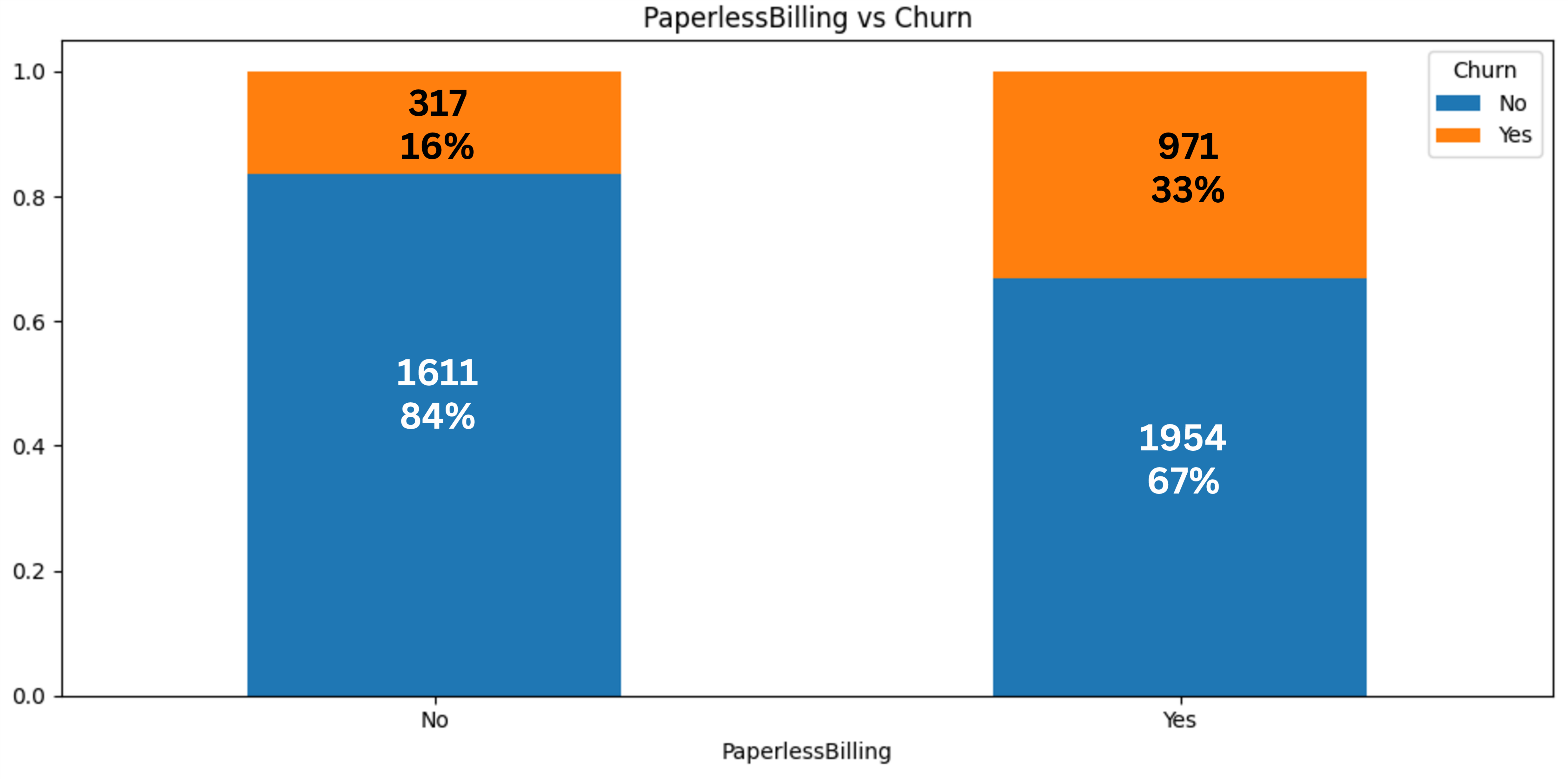



# Month-to-month contract has the largest churn rate



# Customers who churn tend to utilize our paperless billing service

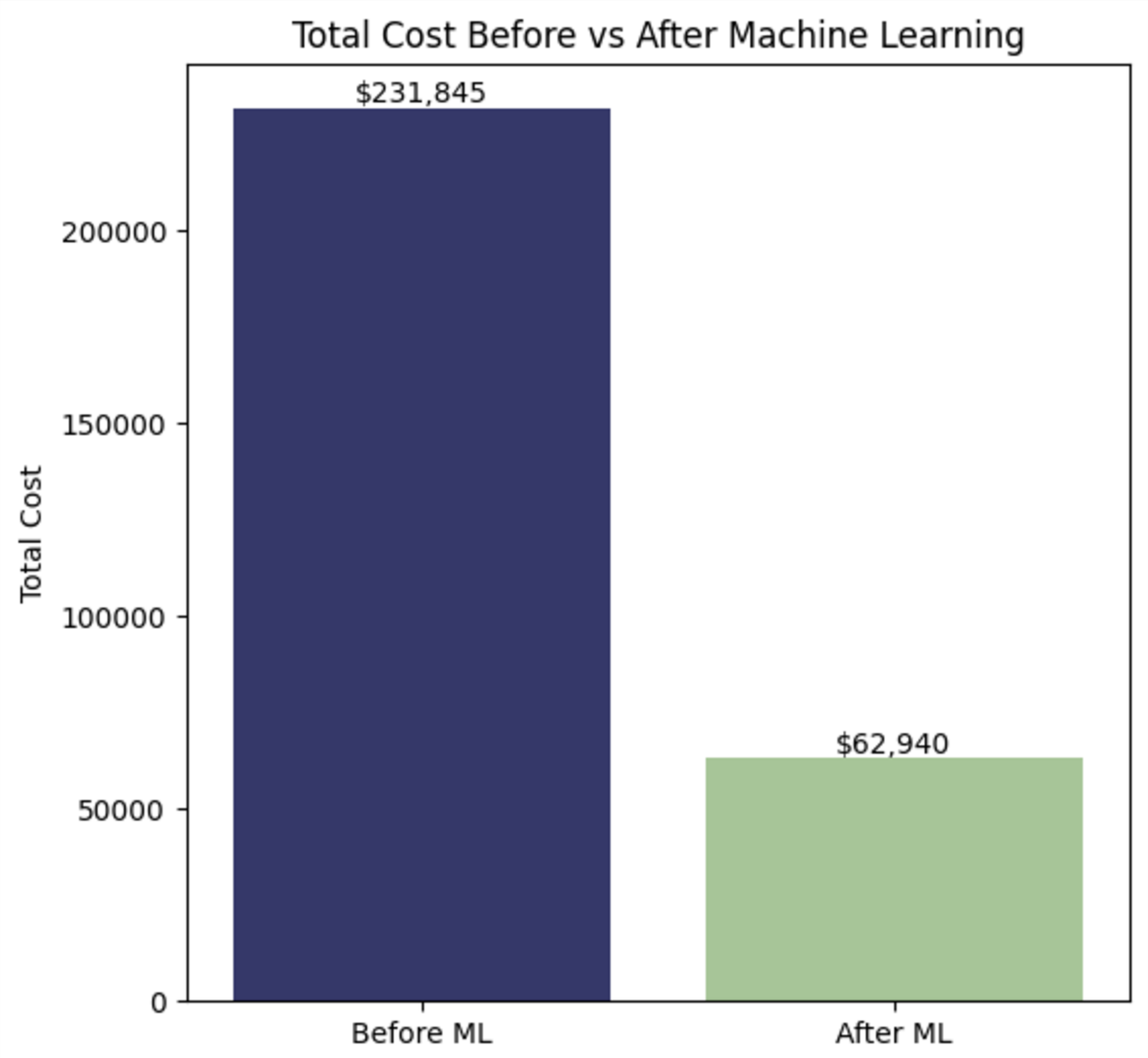


# Impact of Implementing Prediction Model

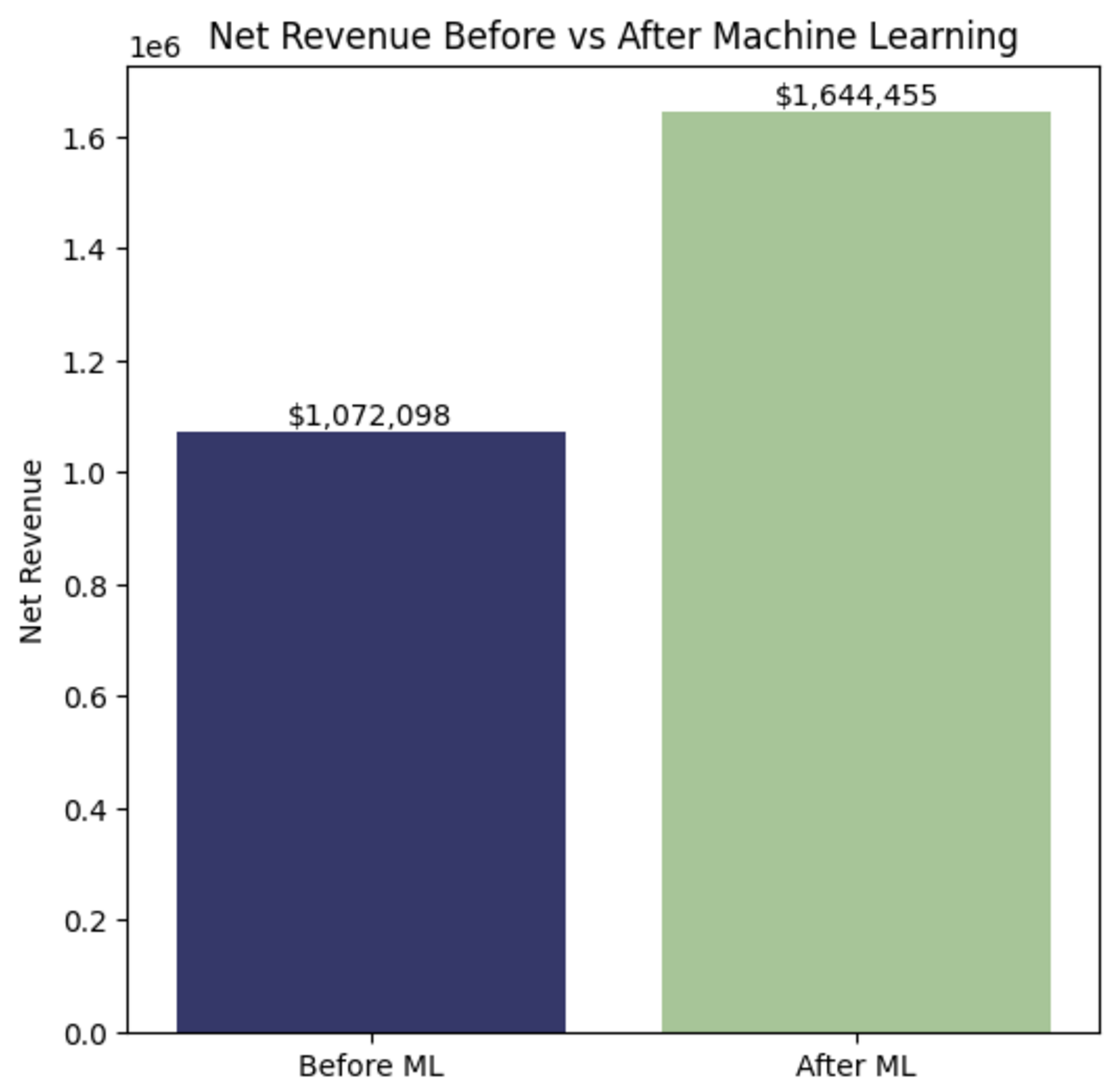
# Using our most recent CAC ($620) and CRC ($125), we find out the following

**Average Revenue per Customer: $63.79**
**Average Customer Lifespan: 32.60**
**Average Customer Lifetime Value: $2079.65**
**Cost of Aquisition: $620**
**Cost of Retention: $125**

# The prediction model helps reduce the customer-related cost our company would accrue by 73%



# It also helps **increase** our **net revenue** by **53%**



# Model Limitations

# Prediction model can only be used if the dataset has the following feature properties

| Feature | Value Limitations |
|---|---|
| `Tenure` | Between 0 to 72 months |
| `MonthlyCharges` | Between 18.8 to 118.65 |
| `Dependents`, `PaperlessBilling` | Yes, No |
| `InternetService` | DSL, Fiber optic, No |
| `OnlineSecurity`, `OnlineBackup`, `DeviceProtection`, `TechSupport` | Yes, No (No and No internet service are treated as the same) |
| `Contract` | Month-to-month, Two year, One year |

# Conclusions

1. Customers with short tenures are more likely to churn compared to those with longer tenures
2. Customers who paid more per month tend to churn more compared to those who paid less
3. Customers with Month-to-Month contract are the ones that are likely to churn
4. Customers who used our fiber optic internet service tend to churn
5. Customers who do not subscribe to our internet-related services are more likely to churn than those who do

With machine learning, we can reduce the total cost up to 73%

- Total Cost Before ML: $231,845
- Total Cost After ML : $62,940

Our net revenue increased 53% with machine learning

- Net Revenue Before ML: $1,072,098
- Net Revenue After ML : $1,644,455

# Recommendations

# To find models that can potentially predict churn better:

1. Increase the amount of customer churn to improve model prediction.
2. Use other resampling methods, including a combination of undersampling and oversampling methods.
3. Explore advanced models using deep learning.
4. Develop customer segmentation models using clustering techniques to identify patterns of customers prone to churn.

# For marketing:

1. Create targeted marketing campaigns for ‘at-risk’ customers.
2. Multichannel marketing to reach customers through their preferred channels.

# For our products:

1. Evaluate the quality of fiber optic service as it has the highest rate of churn.
2. Develop personalized plans based on customer usage history, in this case based on customer tenure and services they subscribe to.

# For sales:

1. Offer additional deals such as monthly charges discounts & bundle deals to customers who have only used the services for 2 years or less
2. Research prices of similar products or services to determine competitive pricing.
3. Offer incentives for long-term contracts such as loyalty rewards to reduce churn associated with month-to-month contracts.

# Thank you!

m.dhany.latief@gmail.com

https://www.linkedin.com/in/dhany-latief-22a674241/

https://public.tableau.com/app/profile/dhan.l/vizzes

# Appendix

# Analytical Approach

- Exploring dataset (EDA)
- Preprocessing data
- Implementing resampling methods: Undersampling and Oversampling methods
- Comparing models based on their f2 scores: Logistic Regression, GradientBoost, SVC
- Hyperparameter tuning
- Choosing model with the highest f2 score: Logistic Regression with IHT resampling method

# Features

| Feature | Description | Impact to Business |
|---|---|---|
| Dependents | Whether the customer has dependents or not | Can influence customer decisions by affecting service needs and preferences, such as the demand for family plans or bundled services |
| Tenure | Number of months the customer has stayed with the company | Often correlates with customer loyalty and retention; longer tenures typically indicate higher customer satisfaction and lower churn rates |
| OnlineSecurity | Whether the customer has online security or not | Impacts the protection of sensitive data, such as personal and financial information |
| OnlineBackup | Whether the customer has online backup or not | This capability fosters customer trust and satisfaction |
| InternetService | Whether the client is subscribed to Internet service | Shows how many of the customers are using the internet service, which can indicate the demand for it |
| DeviceProtection | Whether the client has device protection or not | May enhance customer satisfaction and loyalty in the telecom industry by providing assurance against loss or damage |
| TechSupport | Whether the client has tech support or not | High-quality support can enhance customer loyalty and reduce churn rates |
| Contract | Type of contract according to duration | Longer contract durations typically lead to increased customer loyalty and higher revenue, as subscribers on contracts generate significantly more income than pay-as-you-go customers |
| PaperlessBilling | Bills issued in paperless form | Can lead to cost savings for telecom companies by reducing expenses related to printing and mailing. However, it may also result in consumer confusion and billing complaints |
| MonthlyCharges | Amount of charge for service on monthly bases | Source of revenue for the company, impacting profitability. May also influence customer satisfaction and retention, as consumers often seek competitive pricing and value for their services |
| Churn | Whether the customer churns or not | This feature is to be predicted for the company to be able to make future business decisions |

# Descriptive Statistics

## Numeric Variables

| | tenure | MonthlyCharges |
|---|---|---|
| count | 4930.000000 | 4930.000000 |
| mean | 32.401217 | 64.883032 |
| std | 24.501193 | 29.923960 |
| min | 0.000000 | 18.800000 |
| 25% | 9.000000 | 37.050000 |
| 50% | 29.000000 | 70.350000 |
| 75% | 55.000000 | 89.850000 |
| max | 72.000000 | 118.650000 |

## Categoric Variables

| | count | unique | top | freq |
|---|---|---|---|---|
| Dependents | 4930 | 2 | No | 3446 |
| OnlineSecurity | 4930 | 3 | No | 2445 |
| OnlineBackup | 4930 | 3 | No | 2172 |
| InternetService | 4930 | 3 | Fiber optic | 2172 |
| DeviceProtection | 4930 | 3 | No | 2186 |
| TechSupport | 4930 | 3 | No | 2467 |
| Contract | 4930 | 3 | Month-to-month | 2721 |
| PaperlessBilling | 4930 | 2 | Yes | 2957 |
| Churn | 4930 | 2 | No | 3614 |

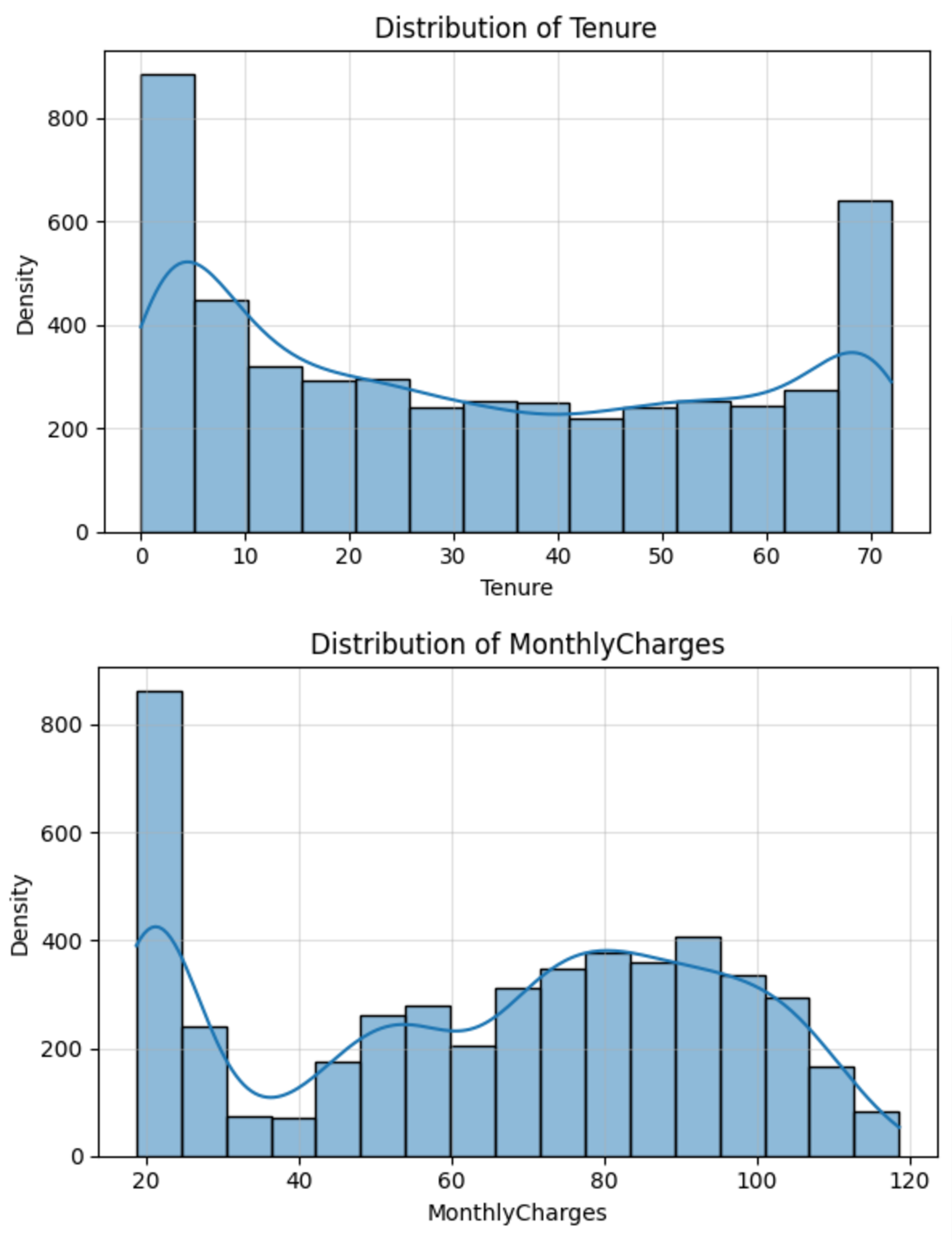
# Data Cleaning

Checking Missing Values: No missing values

Checking duplicated values: 77 duplicated values (1.56% of data) deleted

Checking spelling errors: No spelling errors detected

Changing columns and values: ‘tenure’ -> ‘Tenure’, No internet service -> No




# Numeric Variable Distributions

# Categoric Variable Cardinality

Total Dependents categories: 2
No 0.696476
Yes 0.303524

Total OnlineSecurity categories: 2
No 0.705337
Yes 0.294663

Total OnlineBackup categories: 2
No 0.649289
Yes 0.350711

Total InternetService categories: 3
Fiber optic 0.443849
DSL 0.349475
No 0.206676

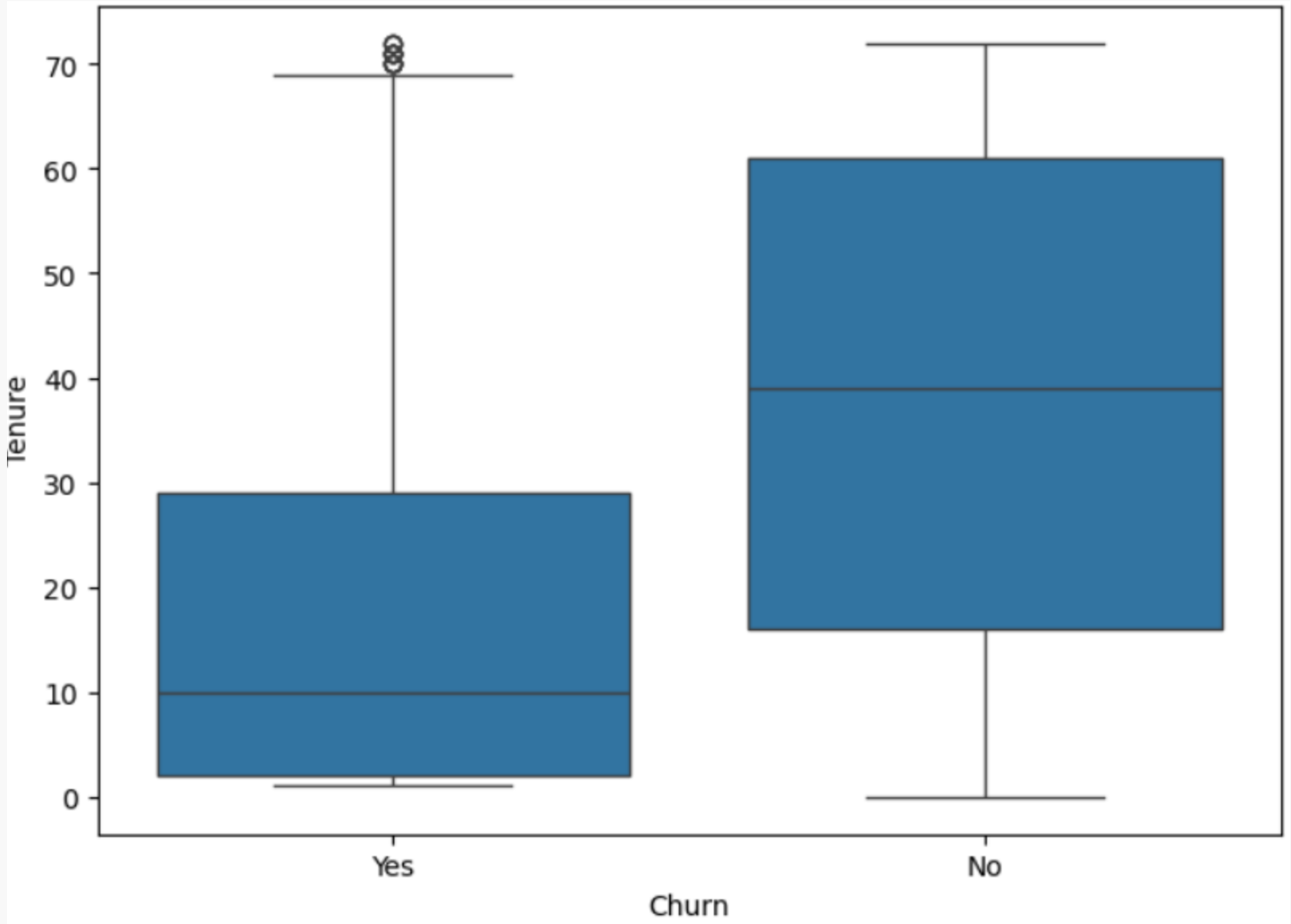
Total DeviceProtection categories: 2
No 0.652174
Yes 0.347826

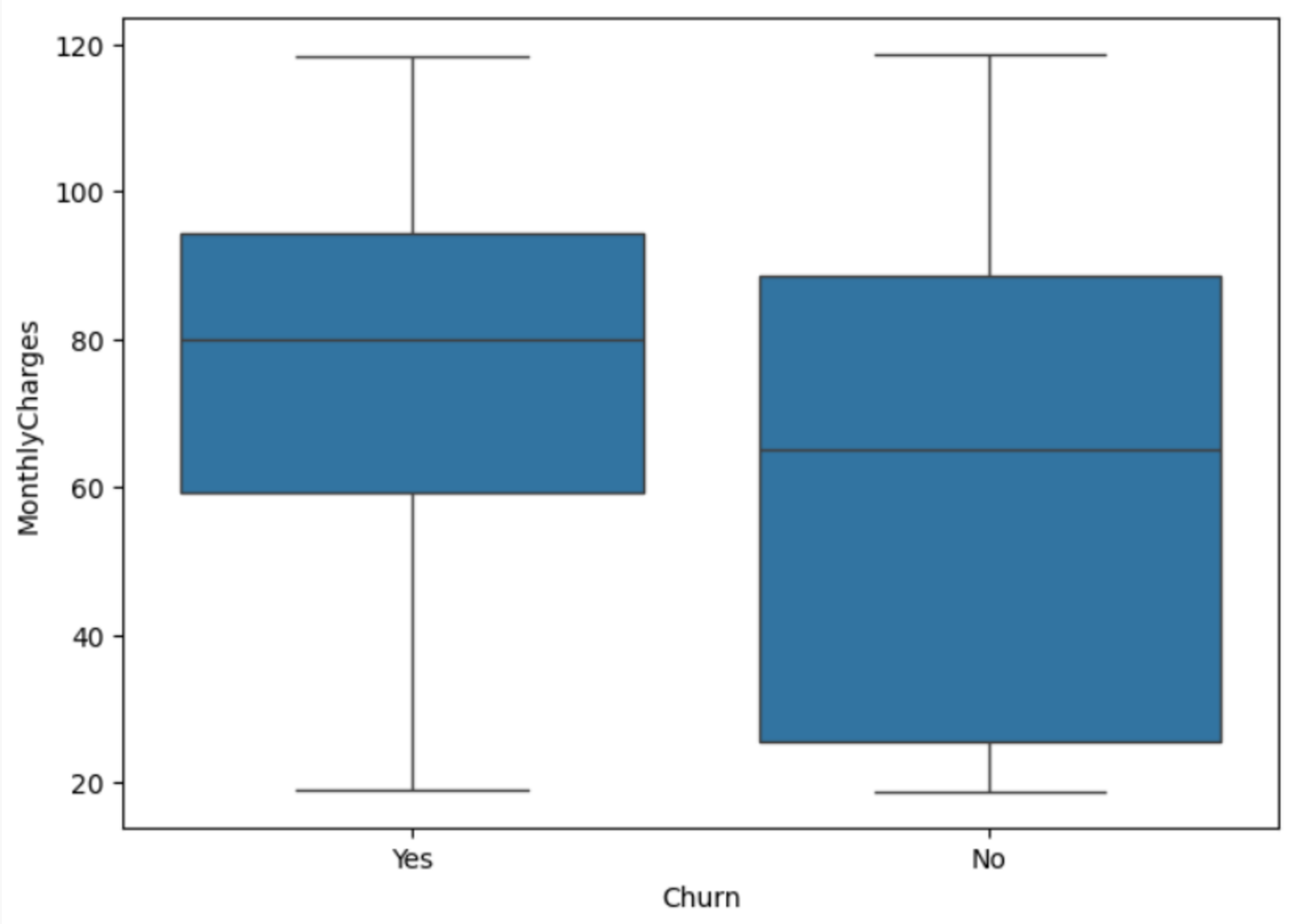
Total TechSupport categories: 2
No 0.70987
Yes 0.29013

Total Contract categories: 3
Month-to-month 0.549145
Two year 0.235524
One year 0.215331

Total PaperlessBilling categories: 2
Yes 0.60272
No 0.39728

Total Churn categories: 2
No 0.734597
Yes 0.265403




# Churn Distributions

# Chosen Models to Compare

| Model | F2 mean cv | F2 sdev | roc_auc mean | roc_auc sdev |
|---|---|---|---|---|
| LGBMClassifier | 0.541733 | 0.036834 | 0.825214 | 0.012974 |
| LogisticRegression | 0.539633 | 0.040555 | 0.837578 | 0.021280 |
| GradientBoosting | 0.531568 | 0.036320 | 0.837970 | 0.016650 |
| XGBClassifier | 0.524202 | 0.040256 | 0.817094 | 0.017435 |
| AdaBoost | 0.519173 | 0.035292 | 0.835596 | 0.021270 |
| KNeighbors | 0.507851 | 0.032757 | 0.771907 | 0.017026 |
| DecisionTree | 0.502120 | 0.030460 | 0.656523 | 0.020933 |
| RandomForest | 0.484781 | 0.032233 | 0.803362 | 0.017868 |
| SVC | 0.476401 | 0.029634 | 0.798761 | 0.024810 |

# Chosen Models to Compare

1. **Logistic Regression with IHT**
Standard Deviation of Cross Validation below 0.02
Increased F2 score after testing (+ 0.01)
High recall = 0.93
ROC AUC = 0.85

2. **GradientBoost with IHT**
Standard Deviation of Cross Validation slightly above 0.02
Slight decrease in F2 score after testing (- 0.01)
High recall = 0.93
ROC AUC = 0.84

3. **SVC with IHT**
Standard Deviation of Cross Validation below 0.02
Very slight decrease in F2 score after testing (- 0.00)
High recall = 0.93
ROC AUC = 0.84

# Logistic Regression Before and After Tuning

Before tuning

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.96 | 0.55 | 0.70 | 571 |
| 1 | 0.43 | 0.94 | 0.59 | 206 |
| accuracy | | | 0.65 | 777 |
| macro avg | 0.70 | 0.74 | 0.65 | 777 |
| weighted avg | 0.82 | 0.65 | 0.67 | 777 |

F2 Score: 0.7581

Best Parameter: {'model__C': 0.5, 'model__max_iter': 50, 'model__penalty': 'l1', 'model__solver': 'liblinear'}

After tuning

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.97 | 0.54 | 0.70 | 571 |
| 1 | 0.43 | 0.95 | 0.59 | 206 |
| accuracy | | | 0.65 | 777 |
| macro avg | 0.70 | 0.75 | 0.64 | 777 |
| weighted avg | 0.82 | 0.65 | 0.67 | 777 |

F2 Score: 0.7623

# Gradient Boost Before and After Tuning

Before tuning

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.96 | 0.54 | 0.69 | 571 |
| 1 | 0.42 | 0.93 | 0.58 | 206 |
| accuracy | | | 0.64 | 777 |
| macro avg | 0.69 | 0.73 | 0.63 | 777 |
| weighted avg | 0.81 | 0.64 | 0.66 | 777 |

F2 Score: 0.7500

Best Parameter: {'model__learning_rate': 0.01, 'model__max_depth': 9, 'model__max_features': None, 'model__min_samples_leaf': 10, 'model__min_samples_split': 2, 'model__n_estimators': 50, 'model__subsample': 0.5}

After tuning

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 0.55 | 0.70 | 571 |
| 1 | 0.43 | 0.93 | 0.58 | 206 |
| accuracy | | | 0.65 | 777 |
| macro avg | 0.69 | 0.74 | 0.64 | 777 |
| weighted avg | 0.81 | 0.65 | 0.67 | 777 |

F2 Score: 0.7514

# SVC Before and After Tuning

Before tuning

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.96 | 0.53 | 0.68 | 571 |
| 1 | 0.42 | 0.93 | 0.58 | 206 |
| accuracy | | | 0.64 | 777 |
| macro avg | 0.69 | 0.73 | 0.63 | 777 |
| weighted avg | 0.81 | 0.64 | 0.65 | 777 |

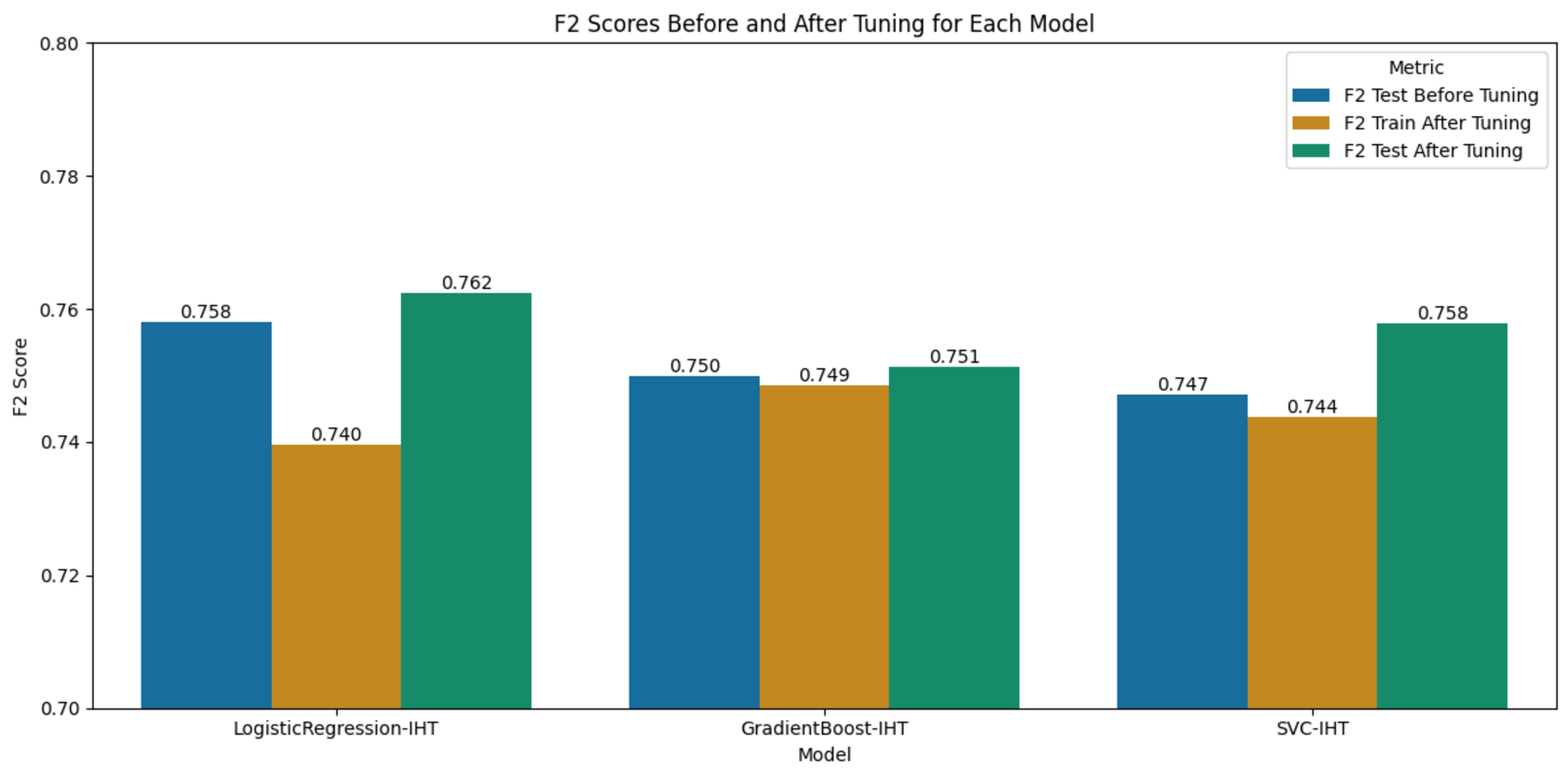
F2 Score: 0.7471

Best Parameter: {'model__C': 0.1, 'model__class_weight': 'balanced', 'model__coef0': 0, 'model__degree': 2, 'model__gamma': 'auto', 'model__kernel': 'poly', 'model__max_iter': -1}

After tuning

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.95 | 0.55 | 0.70 | 571 |
| 1 | 0.43 | 0.93 | 0.58 | 206 |
| accuracy | | | 0.65 | 777 |
| macro avg | 0.69 | 0.74 | 0.64 | 777 |
| weighted avg | 0.81 | 0.65 | 0.67 | 777 |

F2 Score: 0.7578




# Model Comparison

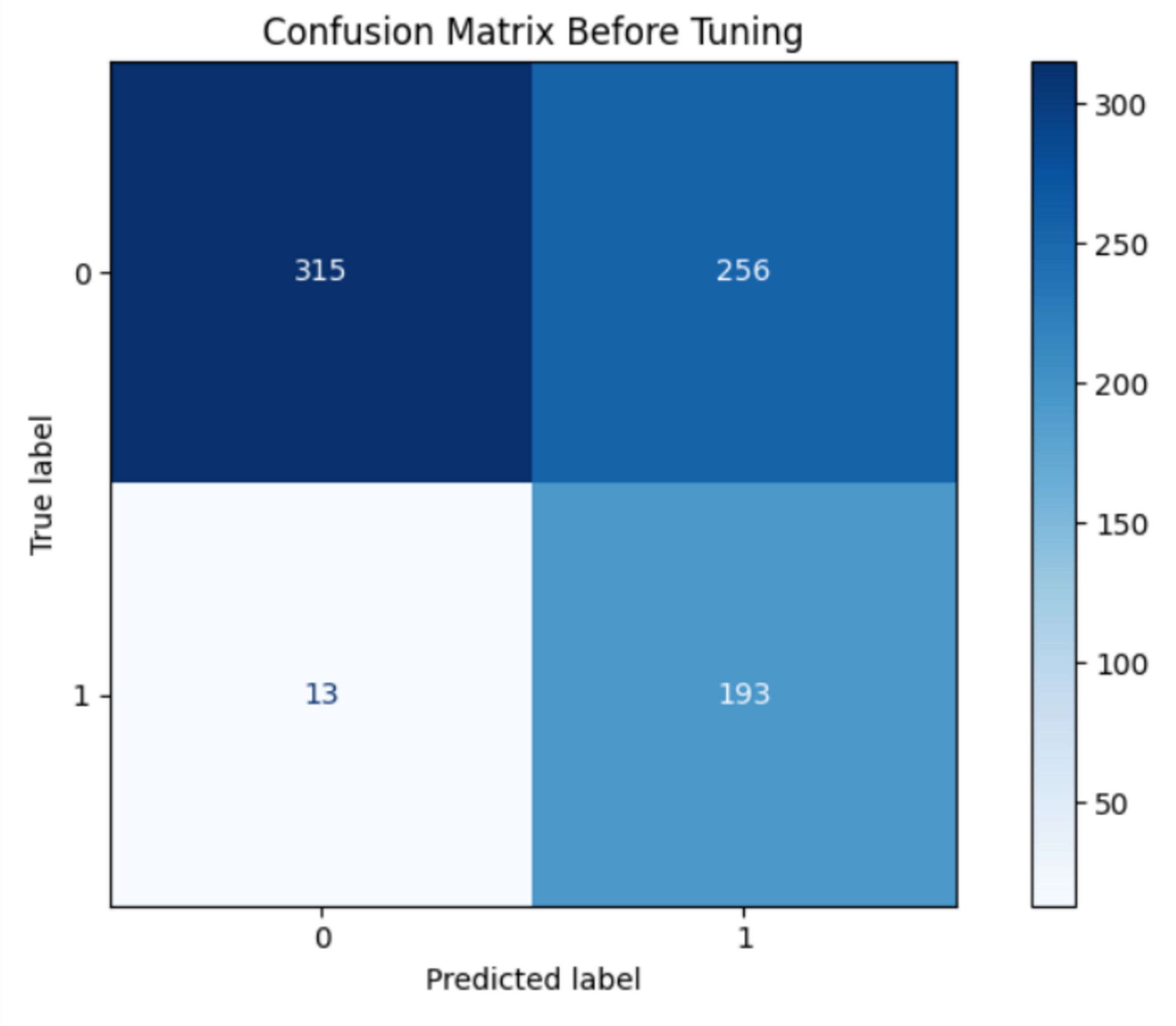
# Best Model Performance

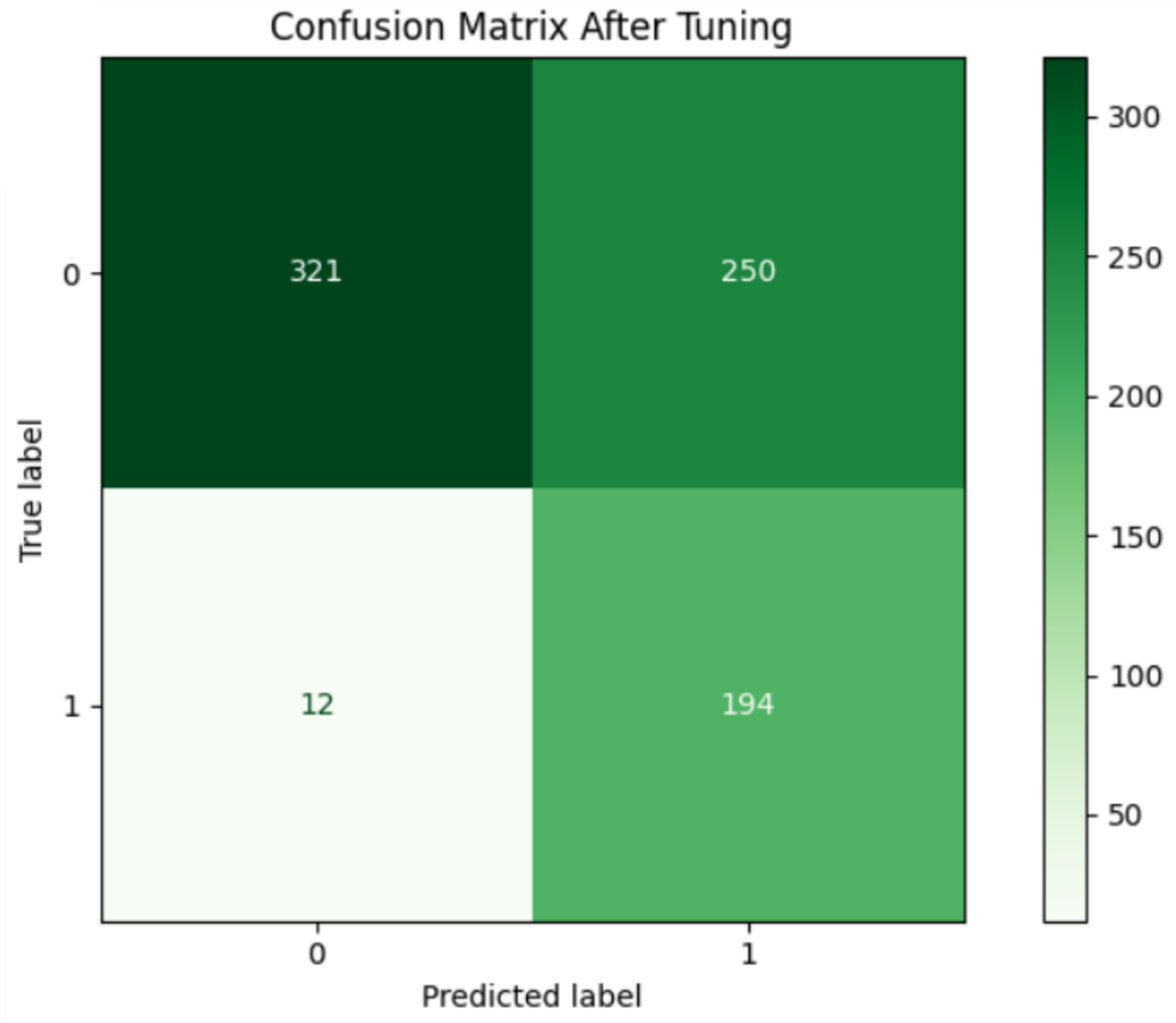
```
F2 Score: 0.7623143080531666
Recall Score: 0.9466019417475728
ROC AUC Score: 0.8500714127828881
              precision    recall  f1-score   support

           0       0.97      0.54      0.70       571
           1       0.43      0.95      0.59       206

    accuracy                           0.65       777
   macro avg       0.70      0.75      0.64       777
weighted avg       0.82      0.65      0.67       777
```

# Best Threshold

| | Threshold | F2 Score |
|---|---|---|
| 52 | 0.53 | 0.764984 |
| 55 | 0.56 | 0.763723 |
| 54 | 0.55 | 0.763116 |
| 49 | 0.50 | 0.762314 |
| 51 | 0.52 | 0.761980 |
| 48 | 0.49 | 0.760530 |
| 50 | 0.51 | 0.760188 |
| 53 | 0.54 | 0.760095 |
| 43 | 0.44 | 0.760031 |
| 47 | 0.48 | 0.759346 |

F2 Score with Optimized Threshold: 0.7649842271293376



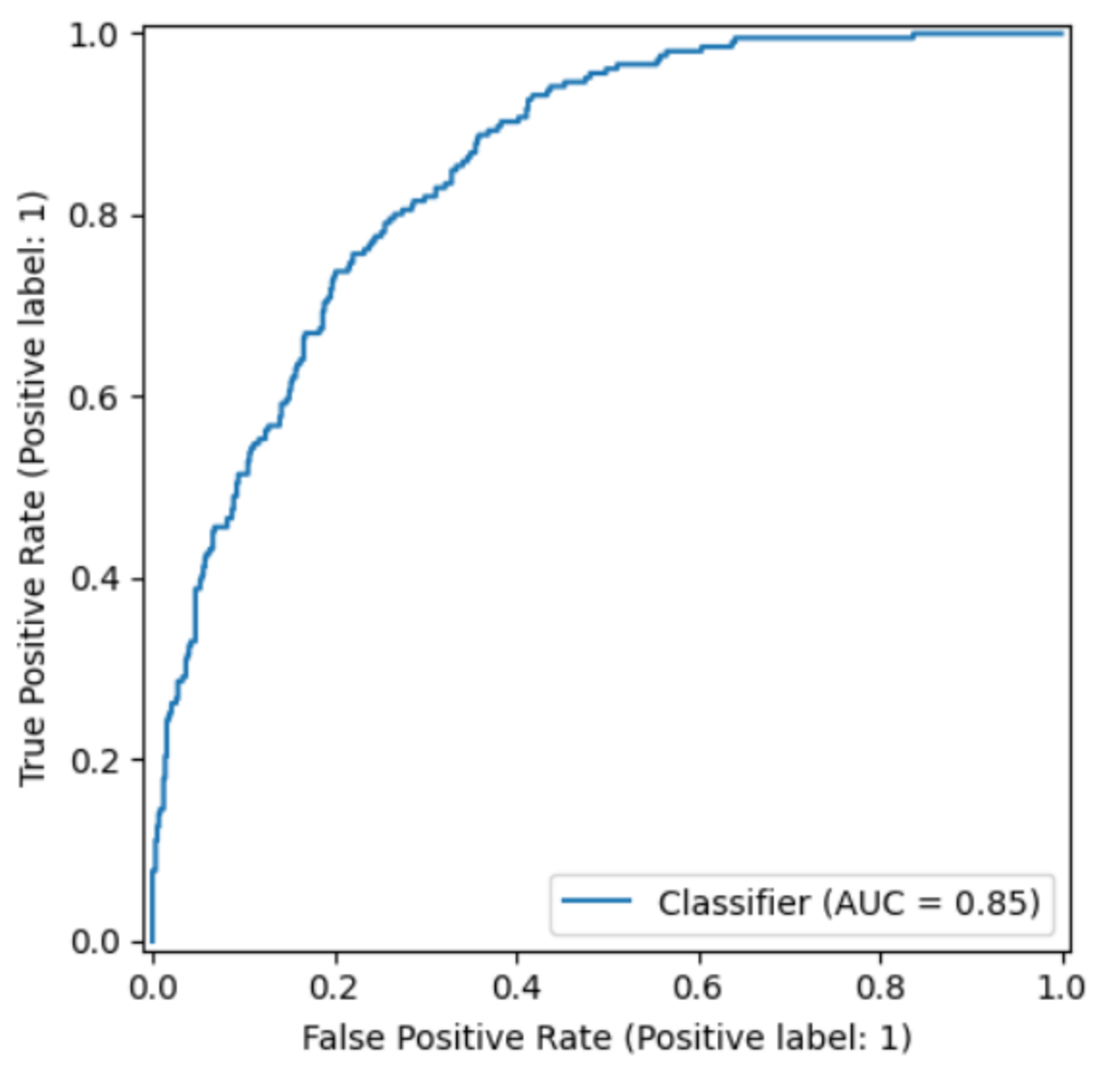




# Confusion Matrix Before and After Tuning

# Confusion Matrix Summary

**Summary of Confusion Matrix**

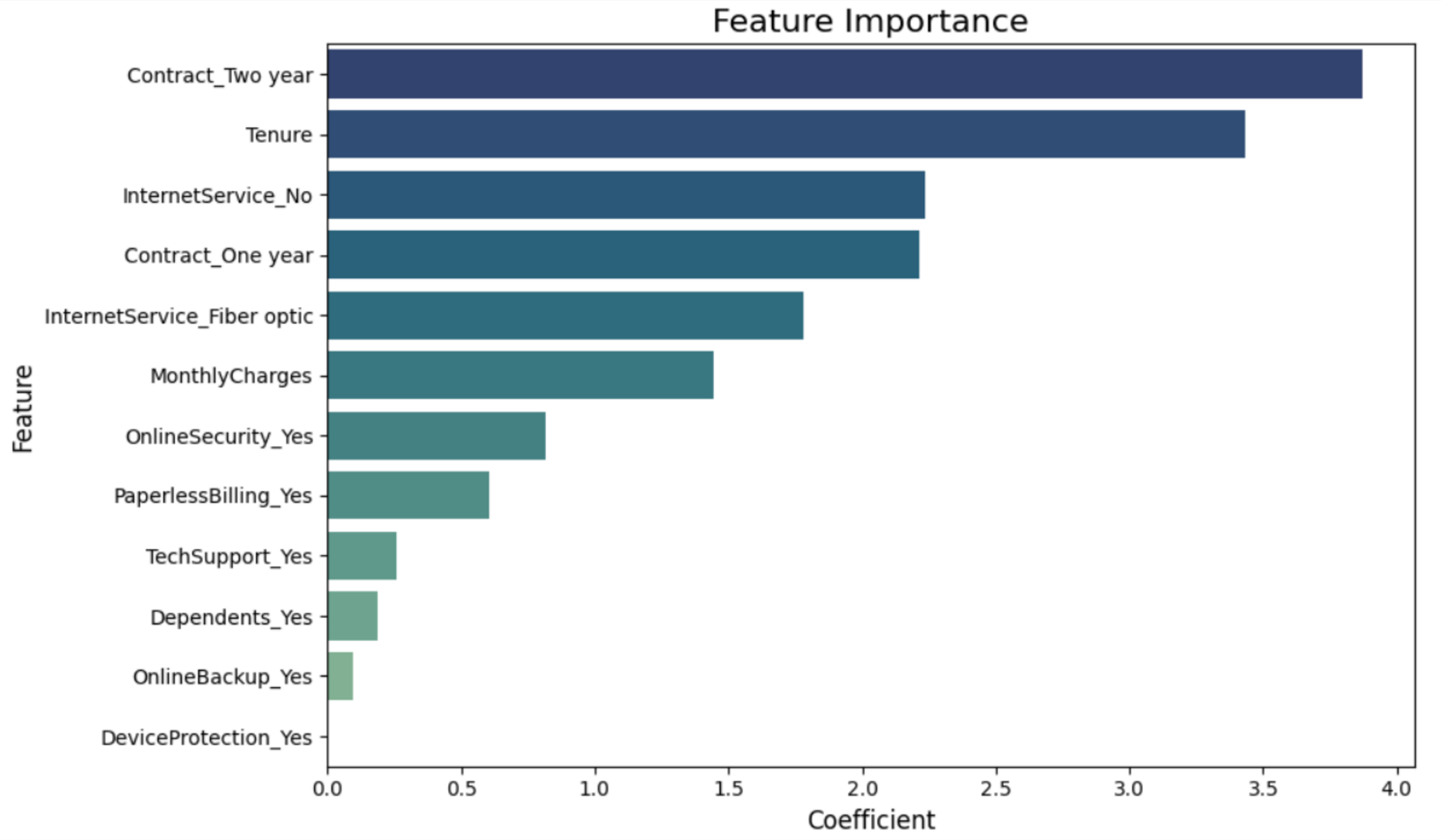
| Label | Description | Total |
|---|---|---|
| True Positive | Customer predicted to churn and actually churns | 194 |
| True Negative | Customer predicted to not churn and actually not churn | 321 |
| False Positive | Customer predicted to churn but actually not churn | 250 |
| False Negative | Customer predicted to not chur but actually churns | 12 |



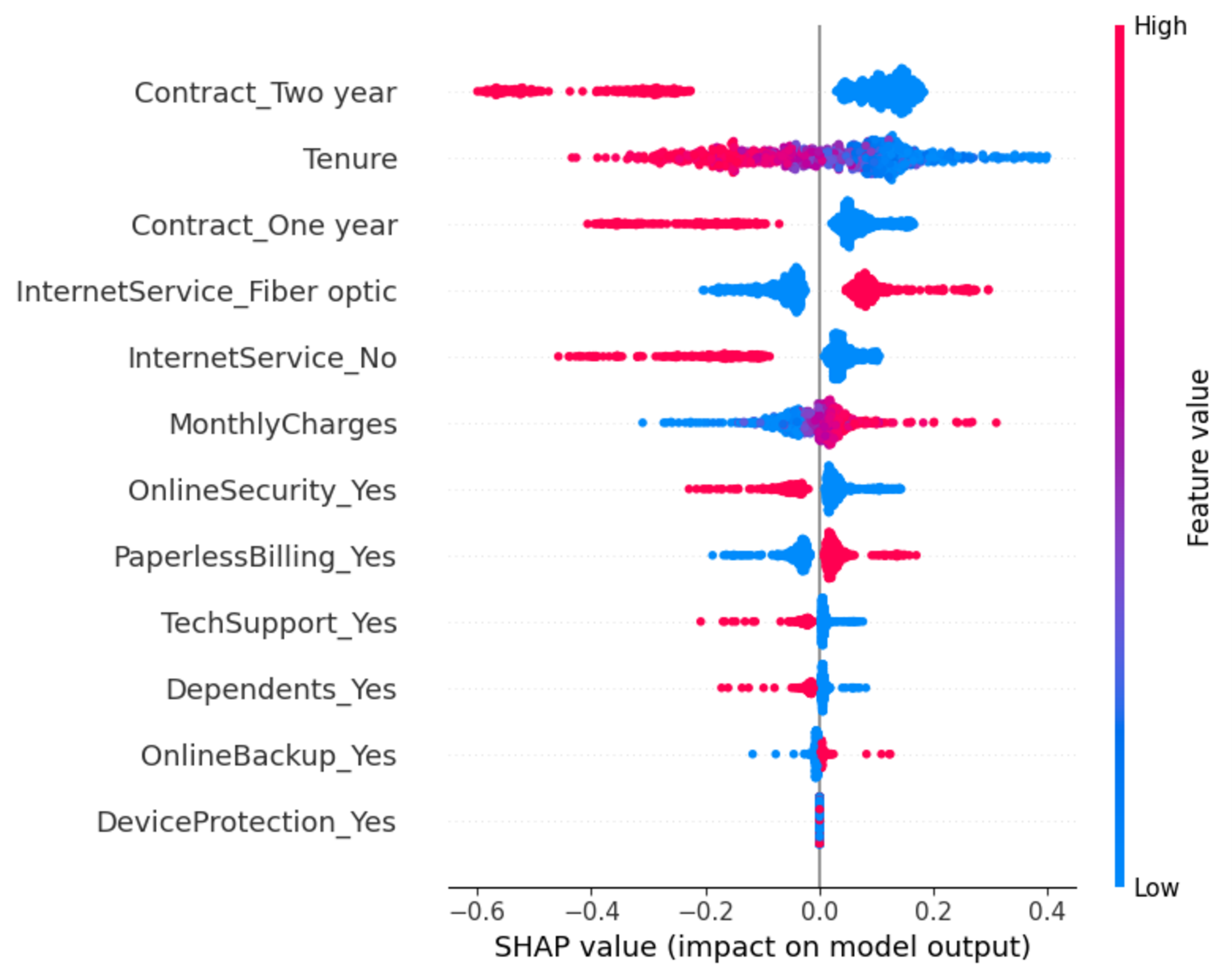

# ROC AUC Curve

# Feature Importance

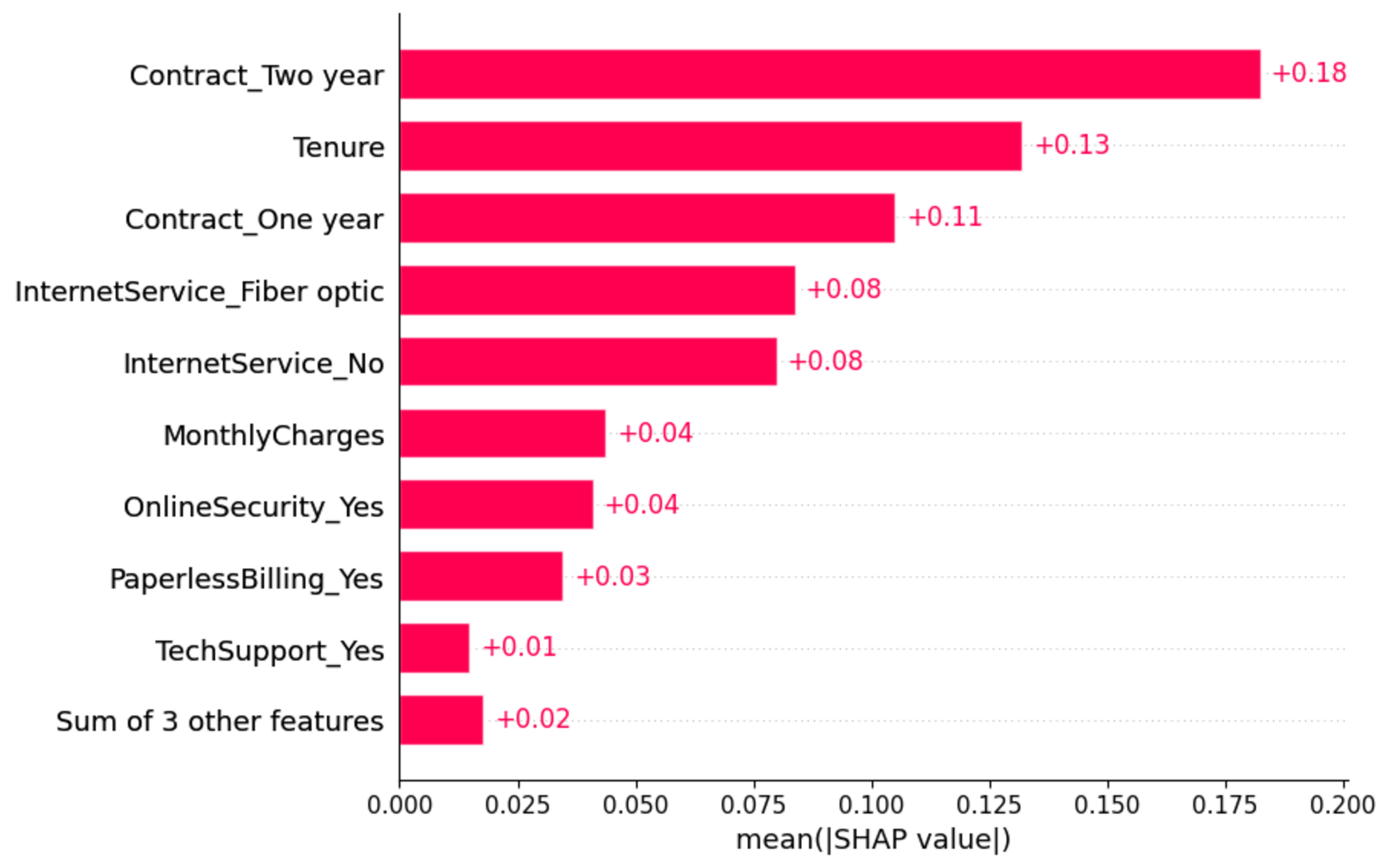
| | Feature | Coefficient | Absolute Coefficient |
|---|---|---|---|
| 8 | Contract_Two year | -3.872845 | 3.872845 |
| 10 | Tenure | -3.430112 | 3.430112 |
| 4 | InternetService_No | -2.237172 | 2.237172 |
| 7 | Contract_One year | -2.216492 | 2.216492 |
| 3 | InternetService_Fiber optic | 1.783289 | 1.783289 |
| 11 | MonthlyCharges | 1.445467 | 1.445467 |
| 1 | OnlineSecurity_Yes | -0.814086 | 0.814086 |
| 9 | PaperlessBilling_Yes | 0.603963 | 0.603963 |
| 6 | TechSupport_Yes | -0.258499 | 0.258499 |
| 0 | Dependents_Yes | -0.188071 | 0.188071 |
| 2 | OnlineBackup_Yes | 0.094313 | 0.094313 |
| 5 | DeviceProtection_Yes | 0.000000 | 0.000000 |



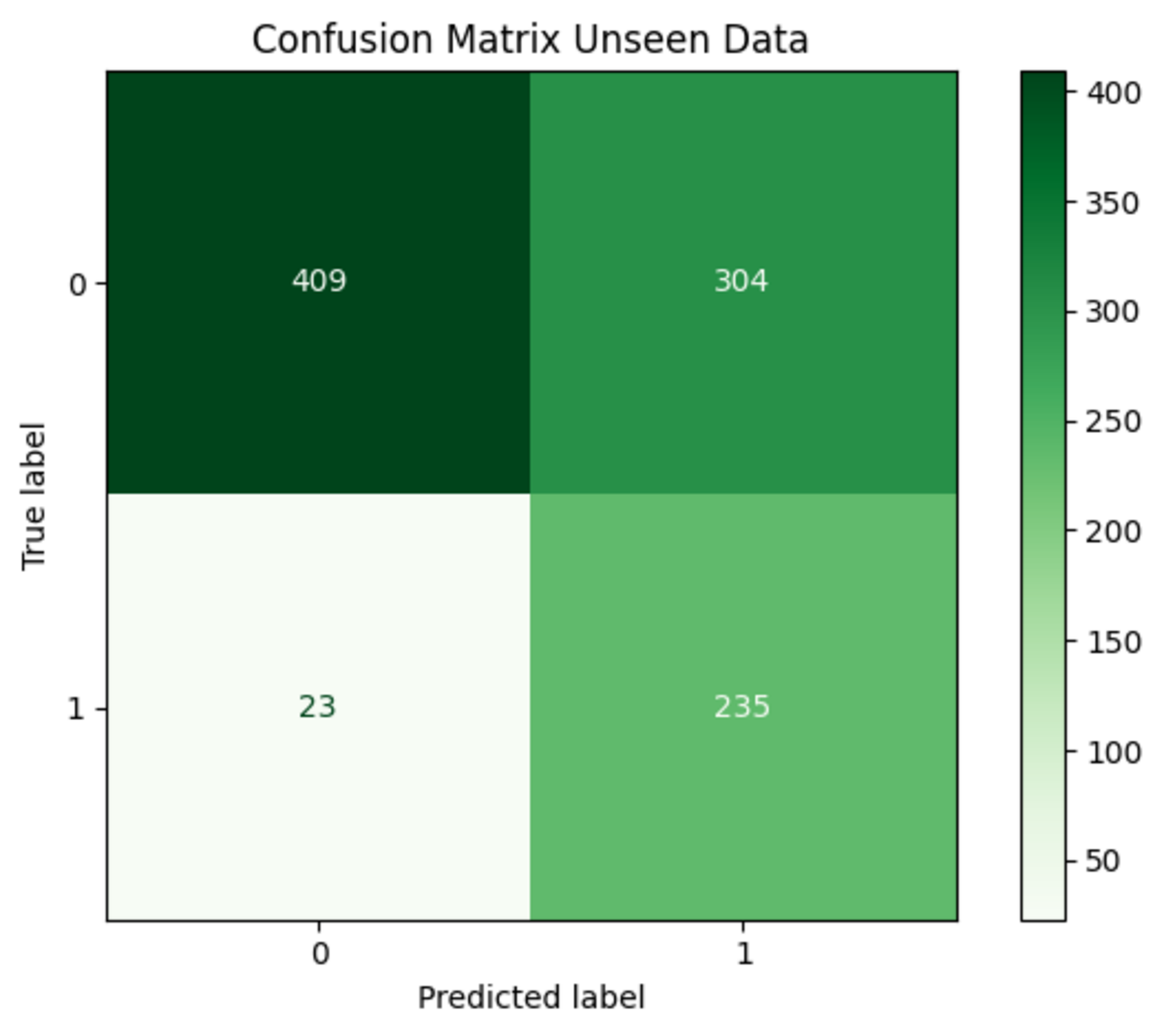

# Feature Importance


Contract_Two year
Tenure
Contract_One year
InternetService_Fiber optic
InternetService_No
MonthlyCharges
OnlineSecurity_Yes
PaperlessBilling_Yes
TechSupport_Yes
Dependents_Yes
OnlineBackup_Yes
DeviceProtection_Yes
−0.6 −0.4 −0.2 0.0 0.2 0.4
SHAP value (impact on model output)
High
Feature value
Low

# SHAP Values


Contract_Two year
Tenure
Contract_One year
InternetService_Fiber optic
InternetService_No
MonthlyCharges
OnlineSecurity_Yes
PaperlessBilling_Yes
TechSupport_Yes
Sum of 3 other features
+0.18
+0.13
+0.11
+0.08
+0.08
+0.04
+0.04
+0.03
+0.01
+0.02
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
mean(|SHAP value|)

# SHAP Values




# Confusion Matrix Business Analysis

# Confusion Matrix Business Analysis

| Label | Description | Total |
|---|---|---|
| True Positive | Customer predicted to churn and actually churns | 235 |
| True Negative | Customer predicted to not churn and actually not churn | 304 |
| False Positive | Customer predicted to churn but actually not churn | 409 |
| False Negative | Customer predicted to not chur but actually churns | 23 |