Project Group #5 - Milestone 3

#### Sirisom Pranivong, Chloe Seo, Kaylee Vo, Kai Yeh

## Data Overview

### Data Description

For this project, we acquired the Telecom Churn data (*Telecom Churn (Cell2cell)*, 2018), which is a cross-sectional, individual-level data set containing 71,047 total observations and 58 features. The dataset was obtained from Kaggle and is the same data used in the paper, A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics by V. Umayaparvathi and K. Iyakutti. The target label is binary, with “Yes” indicating customer churn and “No” indicating retention. The dataset is pre-split into training and testing sets, with the test containing 20,000 observations, representing a 28% test split.

Information from the paper was used to create a data dictionary, containing feature metadata. The features are categorized as follows: customer demography (28 features), usage patterns (11 features), value-added services (6 features), customer care (6 features), billing and payment (3 features), and credit score (2 features). The data schema can be found in Table 7 of the Appendix.

To recap before discussing new material, the data preprocessing steps in MS2 were missingness identification, imputation, basic statistics, scaling, and correcting label imbalance. Not much cleaning was necessary besides imputation. For imputation, we applied a Random Regression strategy for quantitative variables, leaving categorical variables unmodified, as only one categorical feature had missing values. Standardization was selected as the scaling method, since we anticipate future use of regularization techniques. This scaling step was added to a pipeline to streamline future tasks, such as feature engineering and cross-validation. To address label imbalance, we employed random oversampling converting the 0.40 ratio to 1:1.

### Data Summary

For this section, we refrain from discussions on missingness, imbalance and scaling as those topics have already been examined in detail in the prior milestone. Instead, we focus on exploratory data analysis (EDA). Prior to EDA, features were split according to data type: discrete, categorical, continuous, and binary. Grouping variables in this manner facilitated easier chart creation and clearer understanding of each variable's behavior.

#### Discrete Variables

To determine the best features to predict churn we exhaustively plot each feature to identify ones most capable of predicting churn. Starting with the discrete variables, we have a total of 8. Histograms are shown in Figure 1 of the Appendix. The histograms are normalized in order to compare the feature distribution across churn classes. If a feature’s distribution differs between churn classes, there is evidence the feature contains a signal to predict churn.

Of the 8 discrete features, only Months in Service showed evidence of heterogeneity in the churn classes. For those that churned, the distribution follows a Poisson with lambda parameter greater than 1, with a mean around 11 months. As for those that did not churn, the distribution follows a Poisson with lambda less than 1. Recall that for lambda values less than 1, Poisson resembles the exponential distribution while large lambda values approaches normality.

#### Categorical Variables

The difference in distributions for the categorical features is not very pronounced as shown in Figure 2. There is a slight difference in credit ratings, where those that did not churn tended to have higher credit scores. Those that did not churn tended to not be married or have an unknown marital status. All other categorical features had near identical distributions.

#### Continuous Variables

As for continuous features, the ones with the most promise were Monthly Minutes, Total Recurring Charge, and Current Equipment Days. For those that churned, the distribution of Monthly Minutes was highest at zero while those that did not churn the mode was slightly higher than 0. For Total Recurring Charge, those that did not churn had higher values on average. Finally for Current Equipment Days, those that did not churn had lower values compared to those that churned. All distributions can be found in Figure 3a and 3b.

#### Binary Variables

All binary variables were identical across churn classes except for Handset Web Capable. Observations that were handset web capable had a lower proportion of those that churned. As shown by Figure 4 in the top right plot.

## Exploratory Data Analysis

### Deeper Understanding of the Data & Meaningful Insights

Examining “MonthsInService” (Figure 5), the only temporal feature, we identified that churn rate veins around 0.2, then experiences a slight initial drop. Notably around the 10th month mark, the churn rate spikes sharply, suggesting that users may have a contract or commitment that lasts approximately 10 months. Early churn often occurs as promotional rates end or costs increase. Poor signal quality, dropped calls, network congestion, and call setup failures are likely significant contributors as well. Customers may also reconsider their service after short-term contracts expire, especially if they experience frequent service disruptions. This insight indicates a potential need for targeted retention strategies. Following the 10 month spike, the churn rate stabilizes, fluctuating between 0.2 and 0.4, until a period of large fluctuations begins after 50 months. The series shows heavy variation at the end, suggesting potential outliers among long-term users. Churn peaks around 50+ months as many customers reach the end of their contract and may leave for better offers. Aging devices often prompt them to seek upgrades, especially when other providers offer incentives. Additionally, long-term customers who feel overlooked or see better deals are likely to switch.

In terms of class imbalance, we previously identified and addressed the disparity between churned and non-churned users in MS2, which helps reduce bias in trend and pattern identification. Outliers were observed through box plots and distribution plots (Figure 3c), though instead of removing them immediately, we plan to analyze their impact through Cook’s Distance for linear models and use principal component analysis (PCA) to minimize their impact in the non linear models. This approach will allow us to retain valuable data points without impacting the model performance.

Assessing the relationship between variables, strong correlations were observed between features, such as MonthlyRevenue and OverageMinutes (0.786), MonthlyRevenue and MonthlyMinutes (0.710), and Handsets and HandsetModels (0.888). These correlations can bring multicollinearity issues, and we intend to address this by removing one of the correlated features, applying PCA, or using feature selection methods as we build the model. Interestingly, no feature displayed a direct correlation with churn, which suggests that churn might be driven by complex interactions among multiple features rather than a single predictor. This insight will inform the design of more sophisticated models capable of capturing these non-linear relationships.

### Noteworthy Findings

In terms of demographics, the analysis of credit ratings shows a higher churn rate among customers with a high (2-High) credit rating, suggesting a potential link between financial stability and churn. Income Group 6 represents the largest non-zero customer segment, while Income Group 0 shows more varied churn rates, indicating that economic factors may contribute to churn likelihood.

The distributions of continuous variables such as MonthlyRevenue and MonthlyMinutes are heavily right-skewed, with most customers displaying lower revenue and usage, and outliers generating significantly higher revenues. This pattern highlights that a majority of customers are low-cost, low-usage, while a few contribute disproportionately through additional service usage.

Revisiting the multicollinearity issue, we can synthesize new features from correlated pairs. For instance, we can sum OutboundCalls and InboundCalls to create TotalCalls, capturing overall call volume without redundancy. Similarly, TotalPeakOffPeakCalls can be made by summing PeakCallsInOut and OffPeakCallsInOut, representing total call activity, simplifying peak versus off-peak distinctions. For call response behavior, we can subtract UnansweredCalls from ReceivedCalls to create AnsweredCalls, creating a measure of calls answered. Additionally, TotalReceivedUnansweredCalls can be generated by summing ReceivedCalls and UnansweredCalls. This feature would reflect customer engagement and missed opportunities. Consolidating correlated features into new features is a simple and effective approach to reduce multicollinearity and enhance model interpretability.

## Research Questions

With the EDA complete we arrive at the following research questions:

1. Which variables influence churn, and how can these factors be used to effectively predict customer churn?

2. How do customer usage patterns and financial characteristics, such as monthly revenue and overage charges, impact customer churn rates?

3. What role do demographic factors (such as income, credit rating) play in predicting customer churn?

## Baseline Model

For our baseline model, we created a naive model and a logistic regression model and applied them to an unbalanced data set. We decided against using upsampled data since class weights would give us identical results. The features used were the top five correlated features found via the correlation matrices (Figure 6a - 6c). The classification report for the naive model is displayed on Table 8a. The accuracy of the naive model is 0.59, which is expected since the majority class is 60% of the data. The weighted f1-score is 0.59 while the evenly weighted f1-score is 0.50.

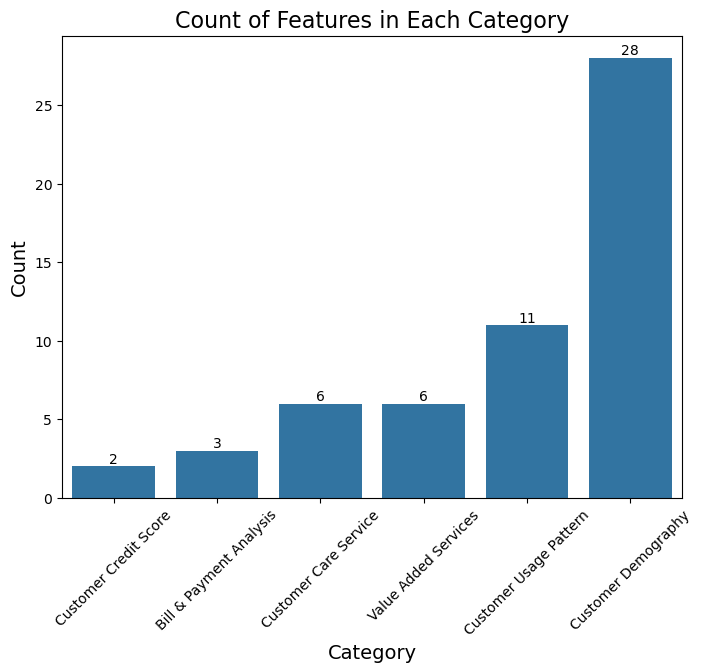
The logistic regression had an accuracy of 0.58 which is slightly worse than the naive model. The weighted f1-score is identical to naive at 0.59. The evenly weighted f1-score is 0.54, slightly higher than the naive model. The ROC-AUC for the model is 0.576 which is only slightly better than naive. These results prompt the need for a better model.

Moving forward, we will iterate on model development and incorporate new features as discussed in Noteworthy Findings. We intend to apply feature selection, generate polynomial terms and will opt for a different model if the decision boundary is too complex for a logistic model.

In MS3, we have conducted an exhaustive investigation of all features, successfully established a baseline, formulated research questions. We have also identified key issues such as multicollinearity and the lack of correlation between features and response. These will serve as our north star going forward.

## Appendix

### Figure 0

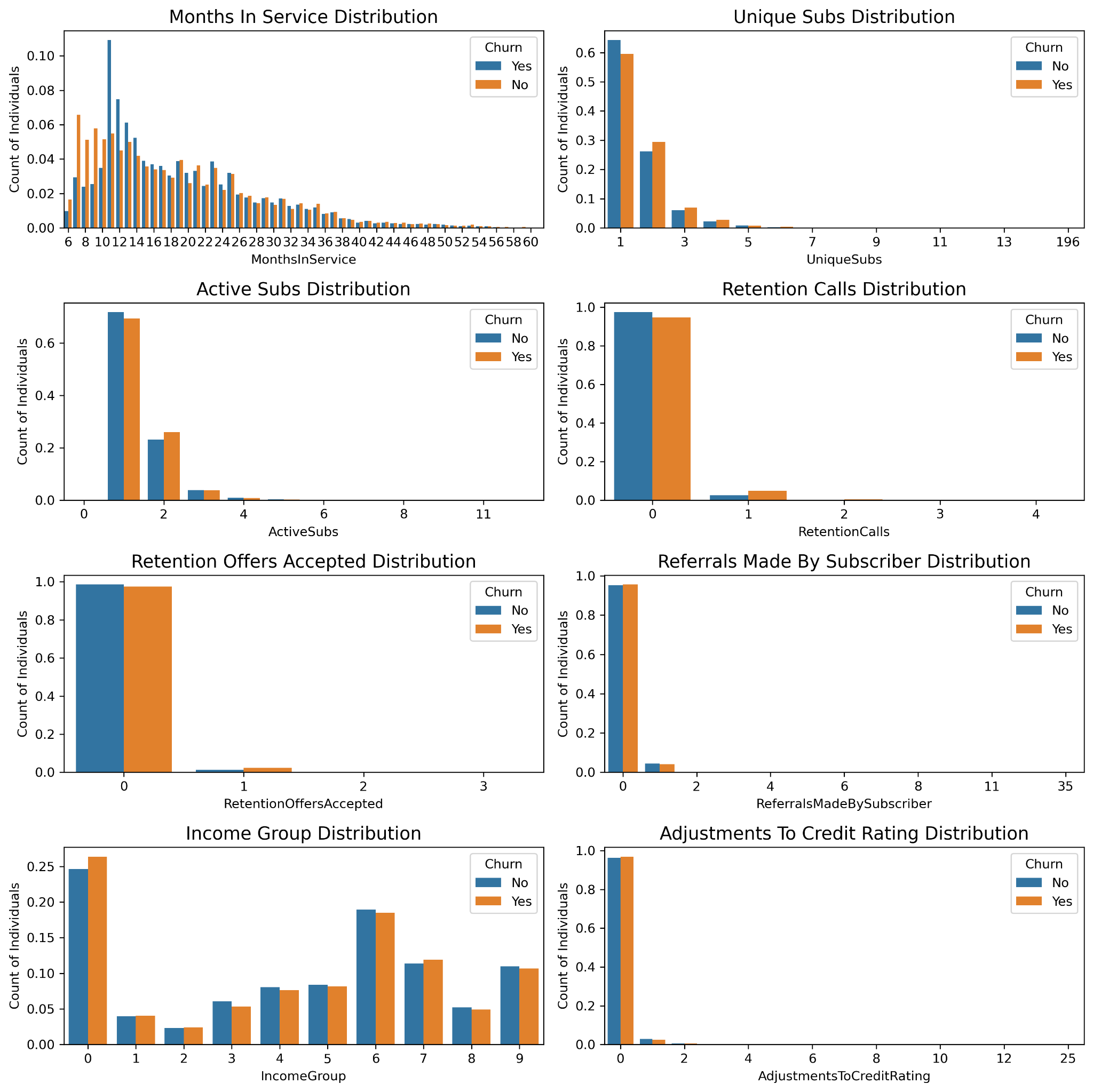


### 

### 

### 

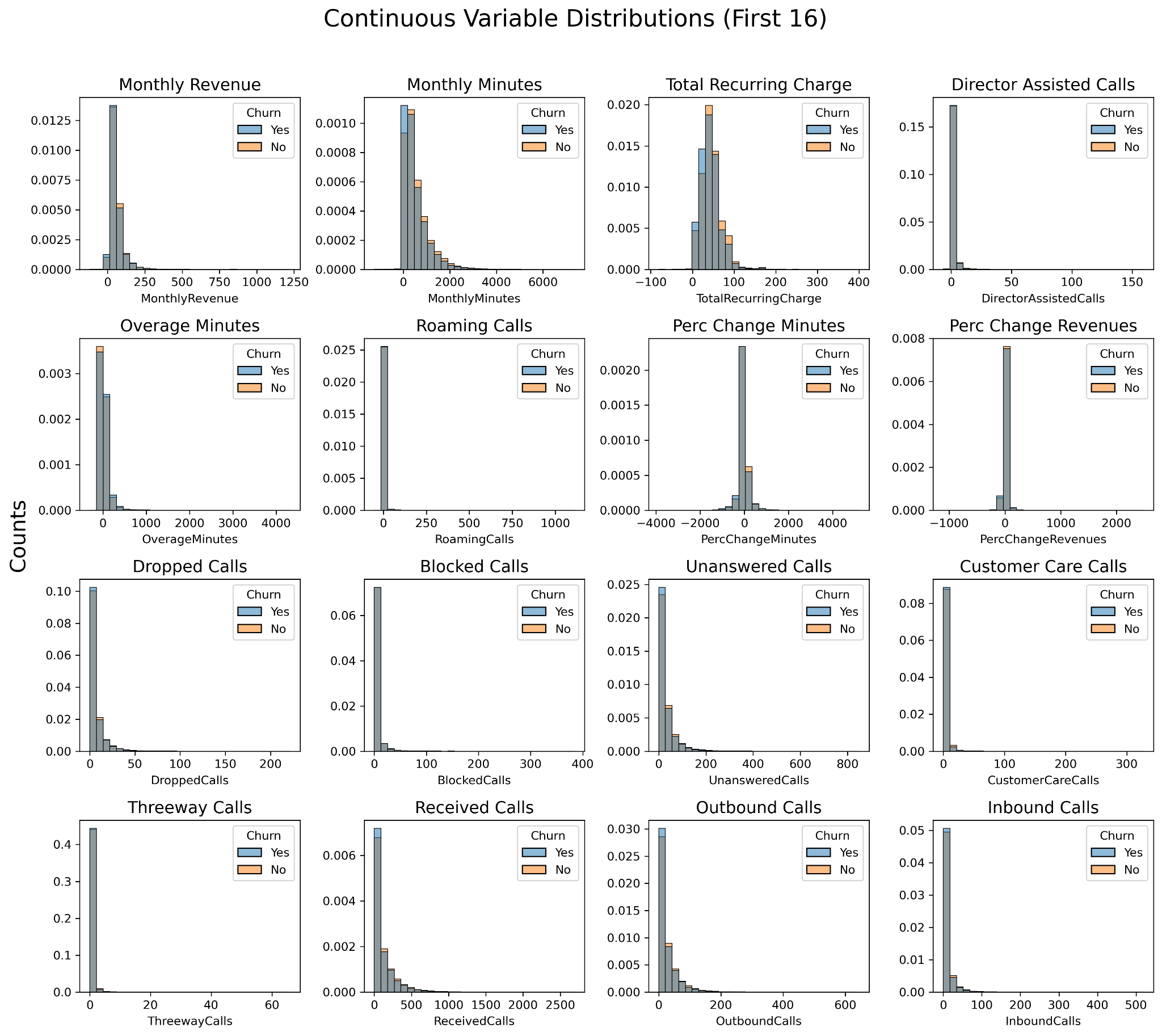
### Figure 1



### Figure 2

### 

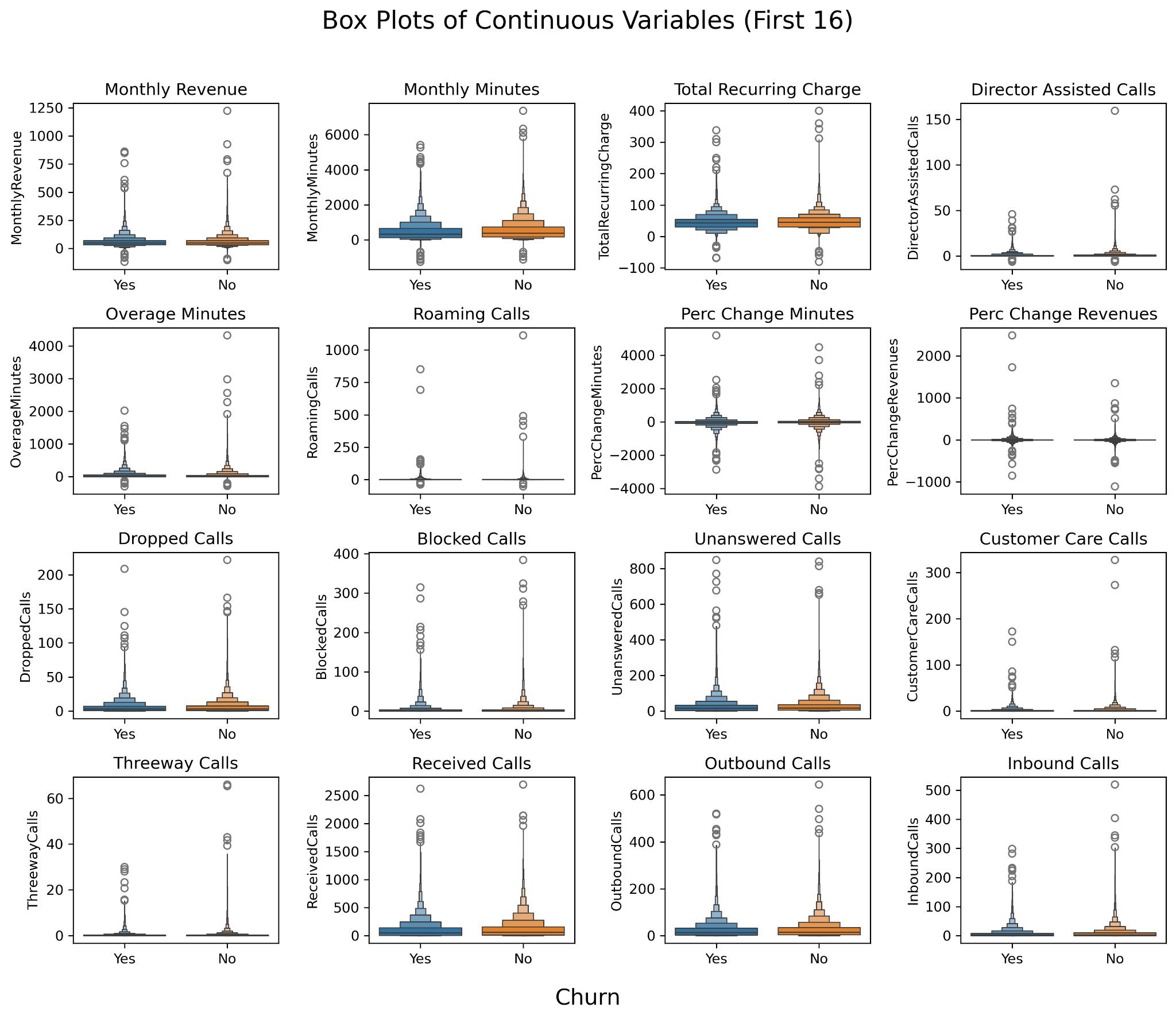
### Figure 3a



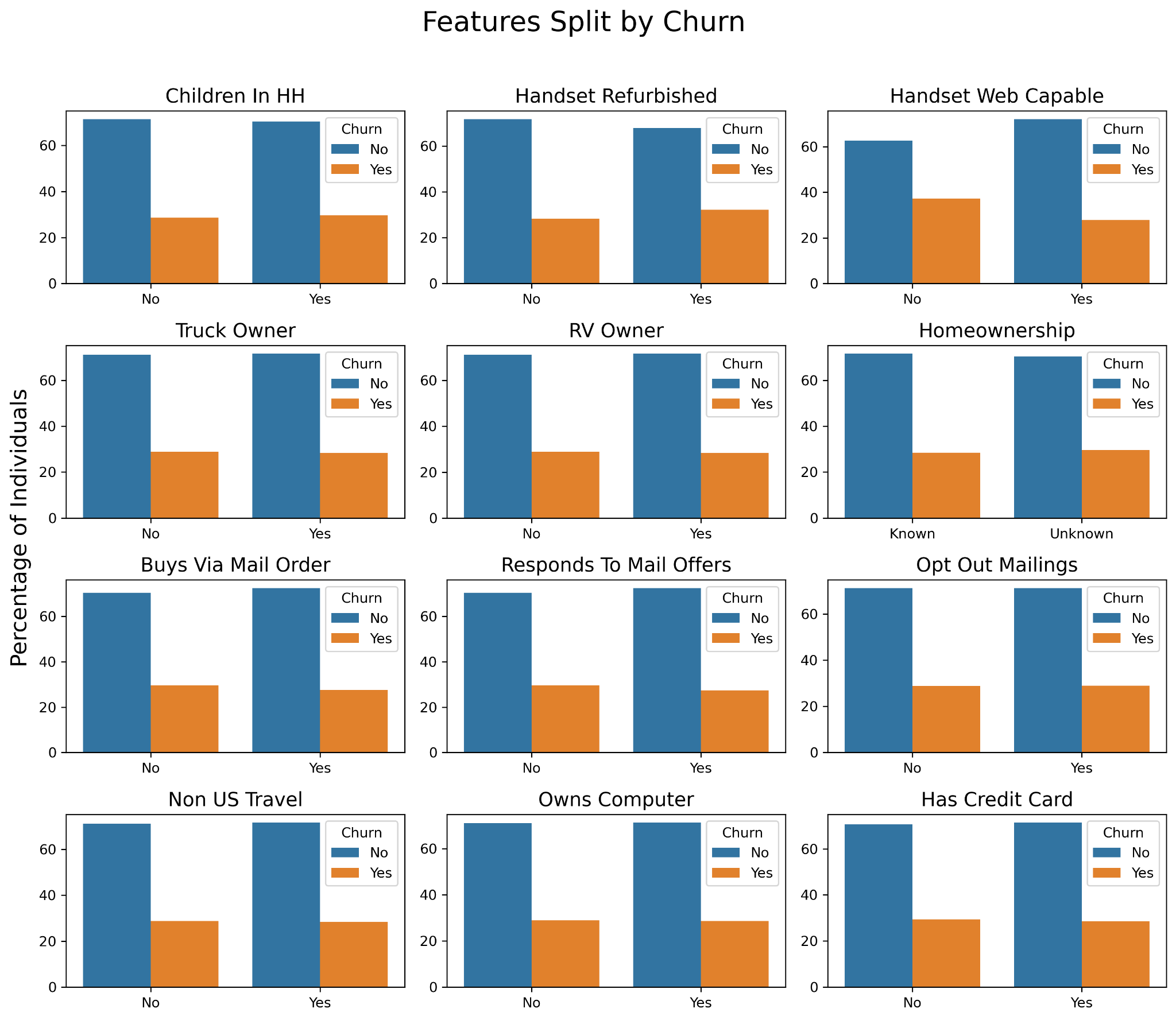
### Figure 3b



### Figure 3c



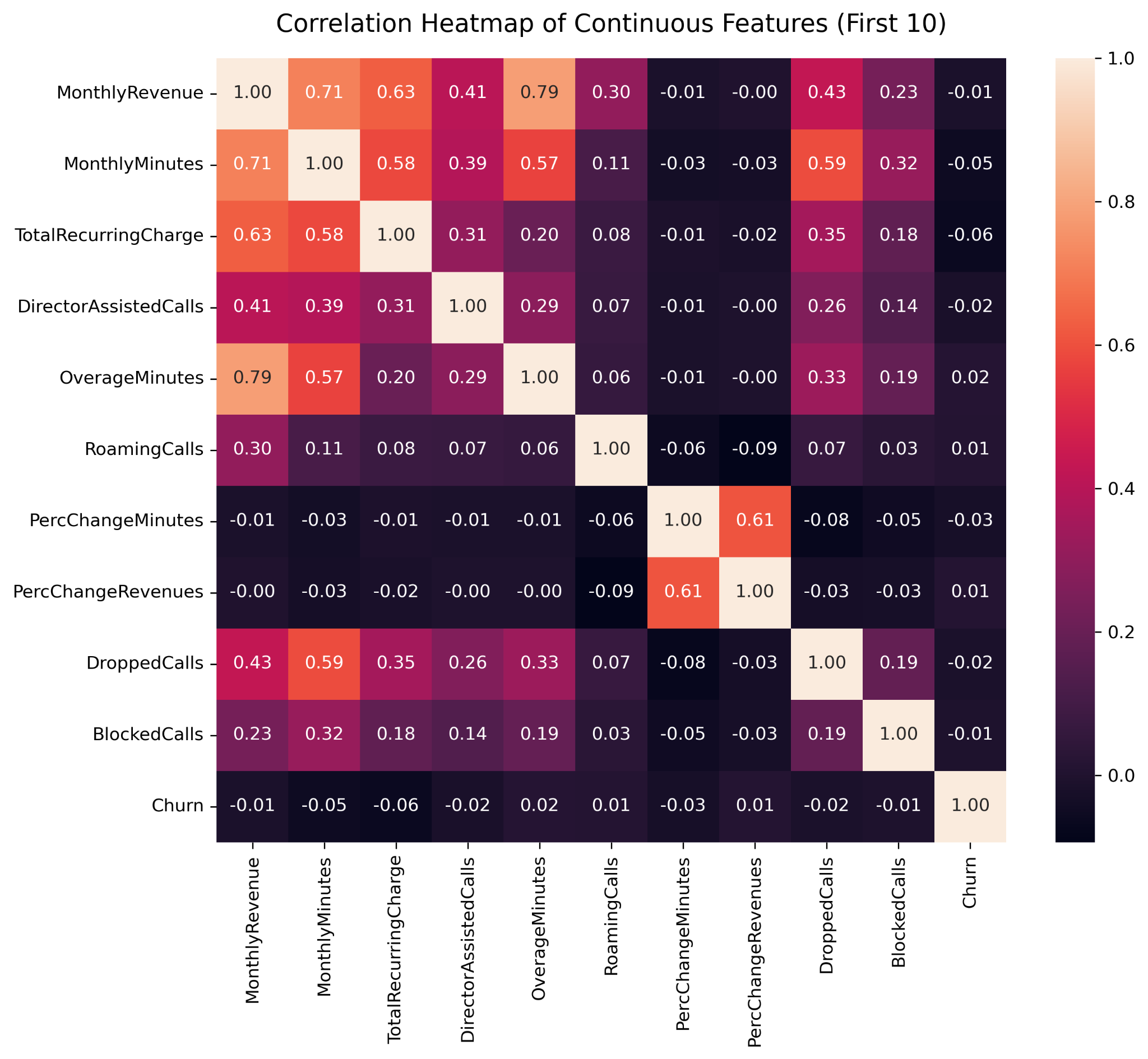
### Figure 4



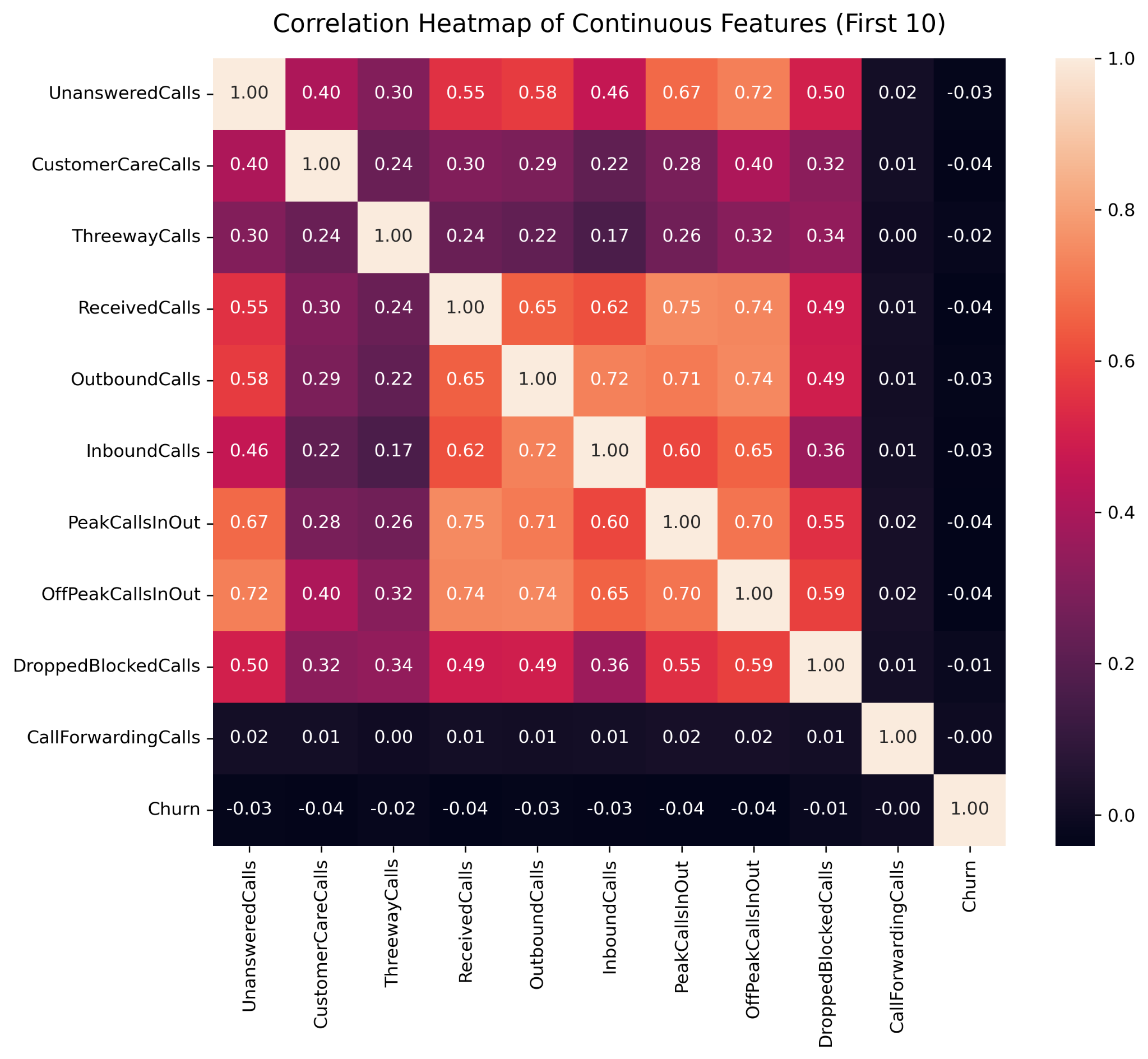
### Figure 5



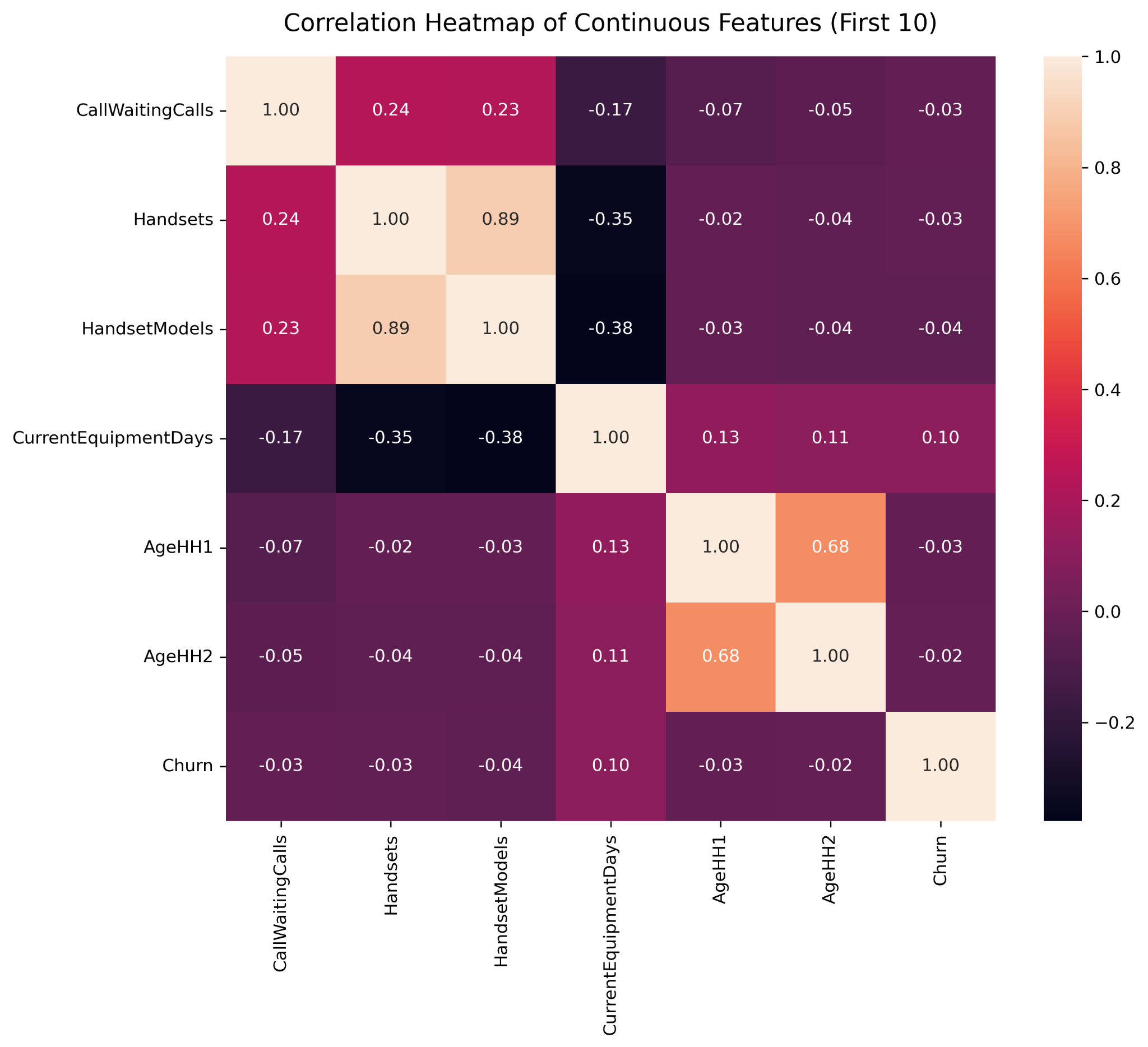
### Figure 6a



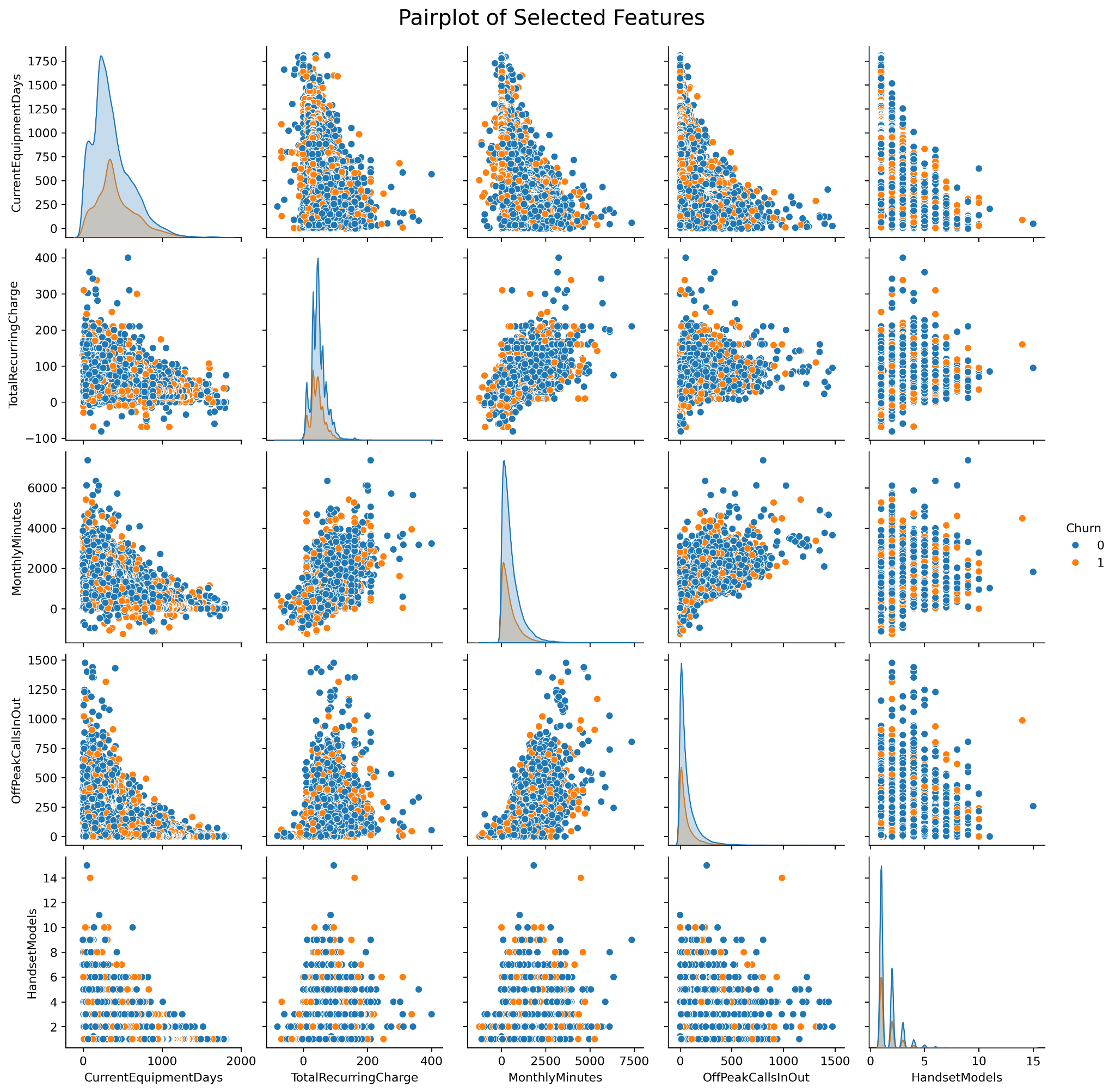
### Figure 6b



### Figure 6c



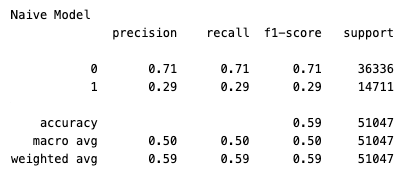
### Figure 7



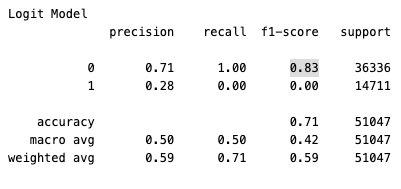
### Table 7 - Data Schema

****

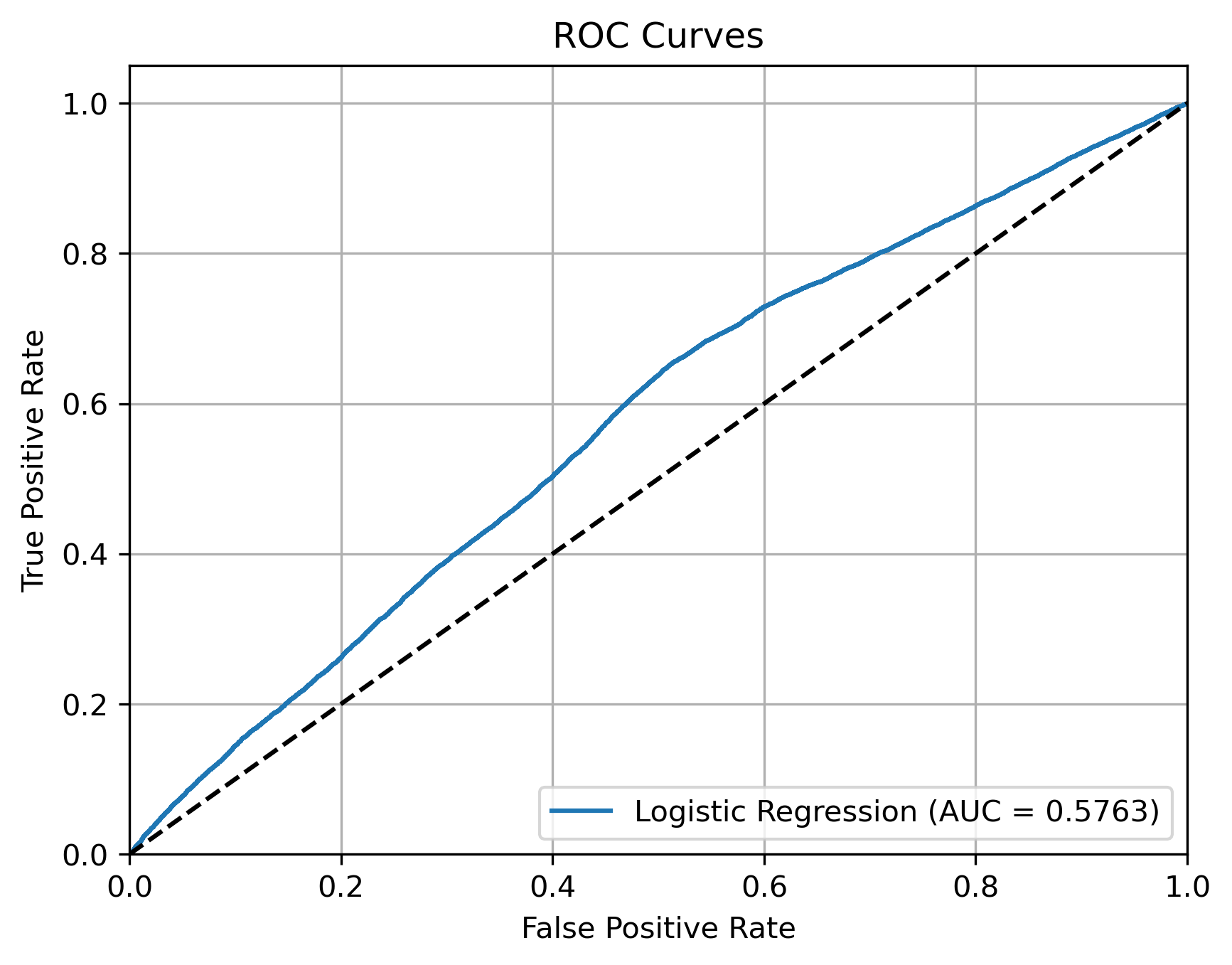
### Table 8a



### Table 8b



### Figure 8



## Sources

Gelman, A., & Hill, J. (n.d.). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Kamaldeep. (2024, October 20). *How to Improve Class Imbalance using Class Weights in Machine Learning?* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/

Kim, M. (2023, December 16). SMOTE: Practical Consideration & Limitations - Minju Kim - Medium. *Medium*. <https://medium.com/@minjukim023/smote-practical-consideration-limitations-f0d926b661a8>

*telecom churn (cell2cell)*. (2018, December 14). Kaggle. https://www.kaggle.com/datasets/jpacse/datasets-for-churn-telecom/data

Umayaparvathi, V., & Iyakutti, K. (2016). A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics. *International Research Journal of Engineering and Technology (IRJET)*, *03*(04). https://www.irjet.net/archives/V3/i4/IRJET-V3I4213.pdf