Telco Customer Churn Analysis

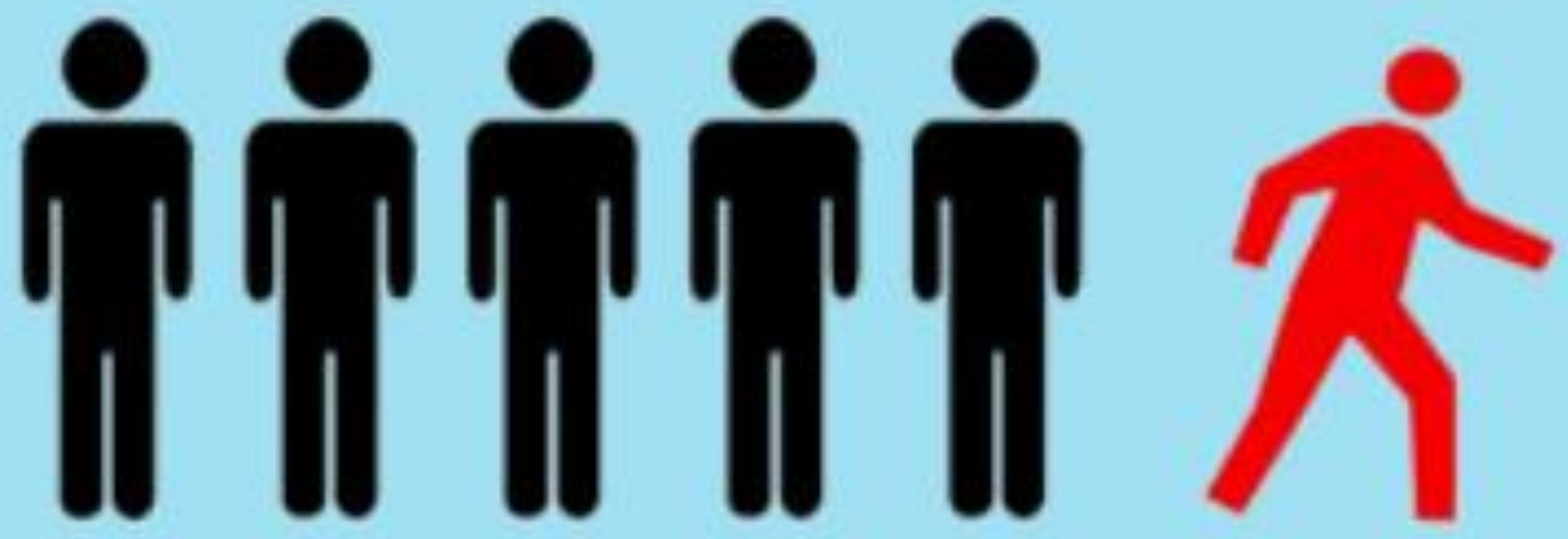
Project member:
Contact no:
Email:

Agenda

1. Problem Statement
2. Objectives of the Study
3. About Data
4. Data Visualization
5. Data Preprocessing
6. Hypothesis Testing
7. Feature Selection
8. Model Training and Evaluation
9. Best Model
10. Limitations, Suggestions and Conclusion

Problem Statement

Did you know that attracting a new customer cost 5 times as much as keeping an existing one?



"Annual Churn Rate of Telco Industry is 31%"

Source: Customergauge.com

- '**Customer churn'** is defined as the process of subscribers (either prepaid or post paid) switching from one service provider.

- With the enormous increase in the number of customers using telephone services, the marketing division for a telco company wants to attract more new customers and avoid contract termination from existing customers (churn rate).

- With proper management of customers, we can minimize the susceptibility to churn and maximize the profitability of the company A mechanism needs to be established to analyze the attributed of profitability.

- Identifying these potential customers early on who may voluntarily churn and providing them retention incentives in form of discounts & combo offers will help the organization to retain those customers and reduce revenue loss. subscribers.

Objectives of the Study

Objective 01

Identify the exiting trends and most appropriate factors associate with the customer prediction

Objective 02

Build a high accurate prediction model for predicting the customers who are likely to churn from the network in near future.

If we can archive these goals, the company can easily identify the probable churn customers and they can push attractive campaigns to prevent customer churn.

About Data

Data Source

Kaggle

www.kaggle.com/blastchar/telco-customer-churn

No of Observations

7043

Dependent Variable

Churn (Yes, No)

Categorical variable

Independent Variables

19 (16 Categorical & 3 numeric)

Which can be classified into 3 groups. Such as Demographical information, account information and service information

Demographical Information

(all variables are categorical)

- Gender (Female, Male)
- SeniorCitizen (0, 1)
- Partner (Yes, No)
- Dependents

Service Information

(all variables are categorical)

- PhoneService (Yes, No)
- MultipleLines (No phone service, No, Yes)
- InternetServices (DSL, Fiber optic, No)
- OnlineSecurity (No internet service, No, Yes)
- OnlineBackup
- DeviceProtection
  TechSupport
- StreamingTV
  StreamingMovies

Account Information

(all variables are categorical except Tenure, MonthlyCharge and TotalCharge)

- tenure (numeric values)
- Contract ((Month-to-Month, One year, Two year)
- PaperlessBilling (Yes, No)
- PaymentMethod (Electronic check, Mailed check,...)
- MontlyCharges (numeric)
- TotalCharges (numeric)

ROADMAP

Collecting Data

Choose the telco customer churn data using Kaggle.

2022/01/26 – 2022/01/27

EDA

Data Visualization

2022/01/28 – 2022/02/01

Data Preprocessing

Cleaning the dataset and resampling

2022/01/29 – 2022/02/01

Hypothesis Testing

Test the results of an experiment to see if we have meaningful results

2022/02/02 – 2022/02/03

Feature Selection

Selecting the best features for the analysis

2022/02/04

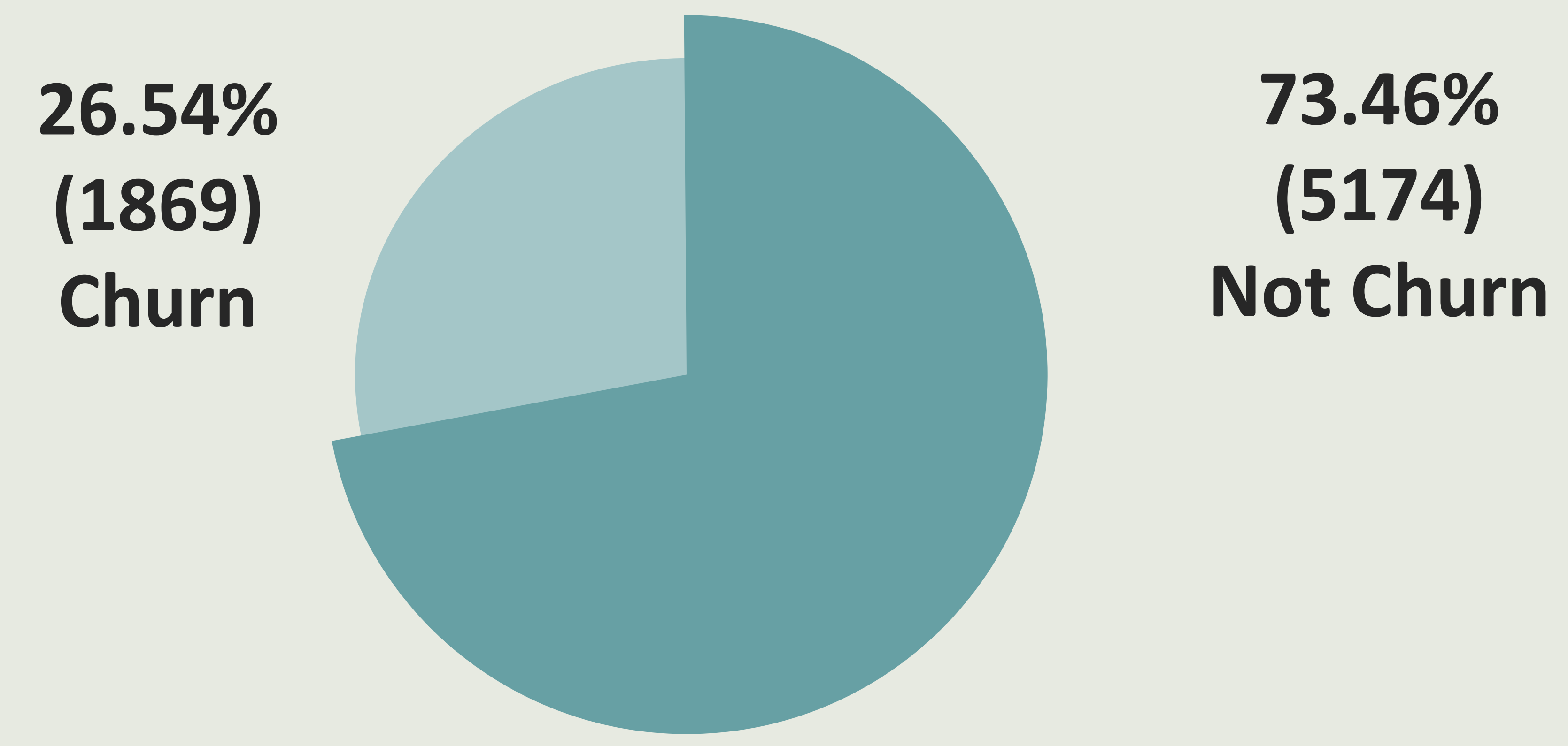
Model Training & Evaluation

2022/02/04 – 2022/02/06

Microsoft Excel, Orange and R Studio will be used for the Analysis

Churn Status

Churn

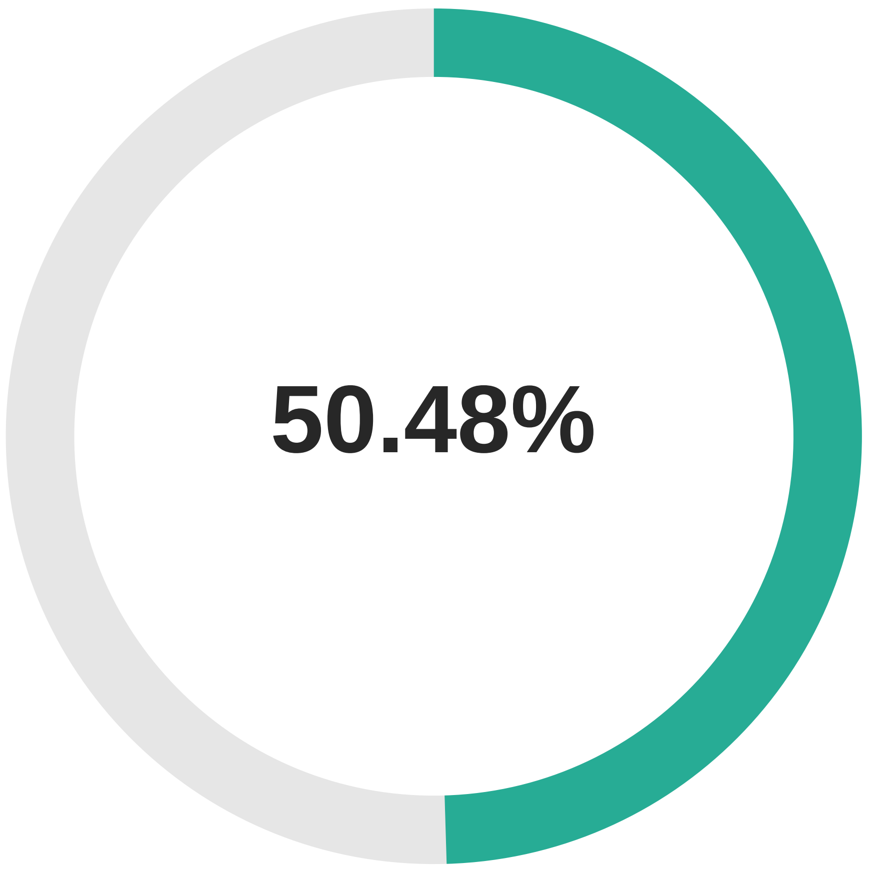


- The response variable of the analysis is churn status of the customer.
- From the sample of 7032 of the customers about 1896 of the customers have churned from the network, which makes it about 26% of the customers have churned.
- The churn percentage is considerable high; therefore, the objective of the study is also clear, that the factors associated with the customer churn should be identified to prevent the customer churn.

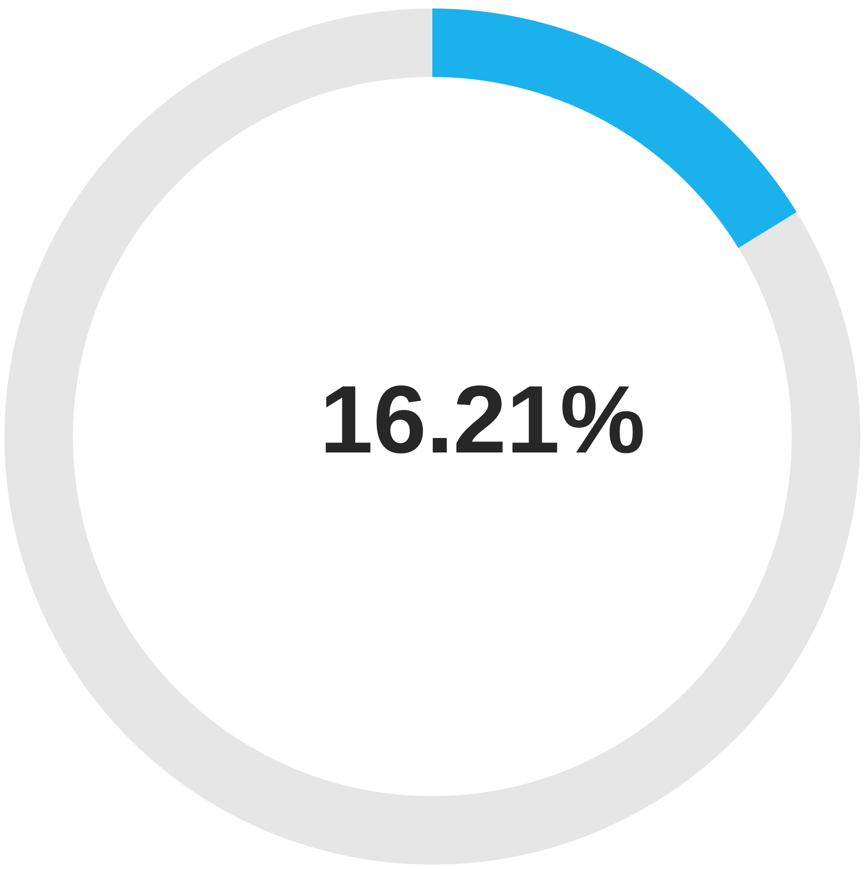
Customer Characteristics

Most of the customers are,

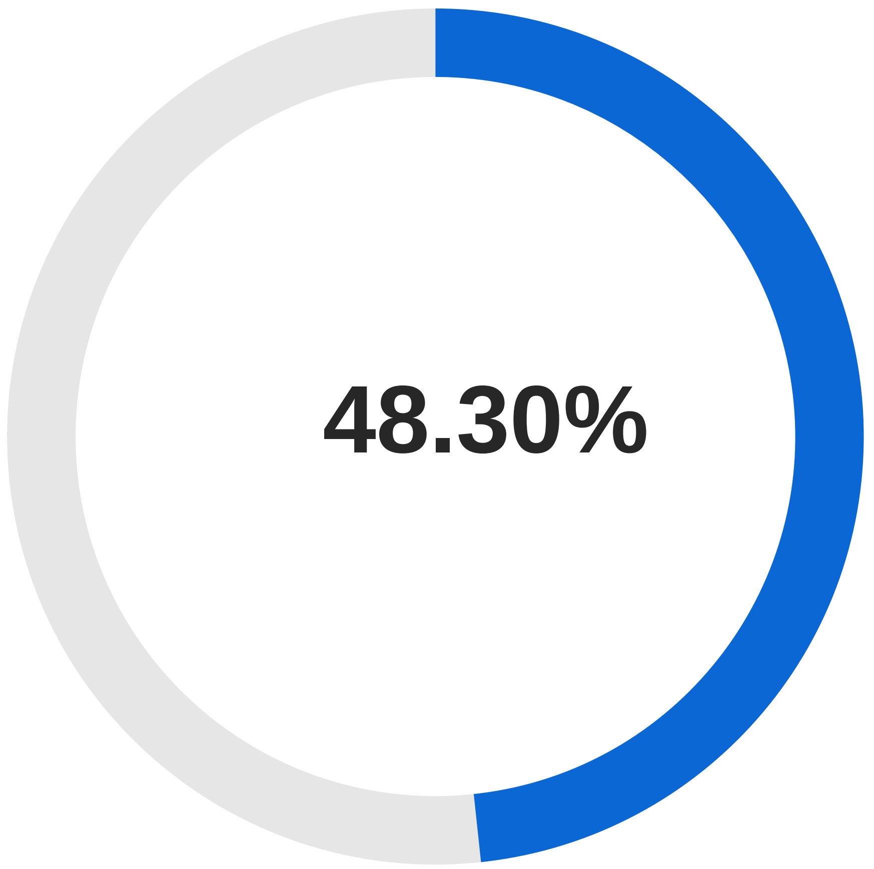
- **Male**
- **Not a senior citizen**
- **Do not a partner and a dependent person**



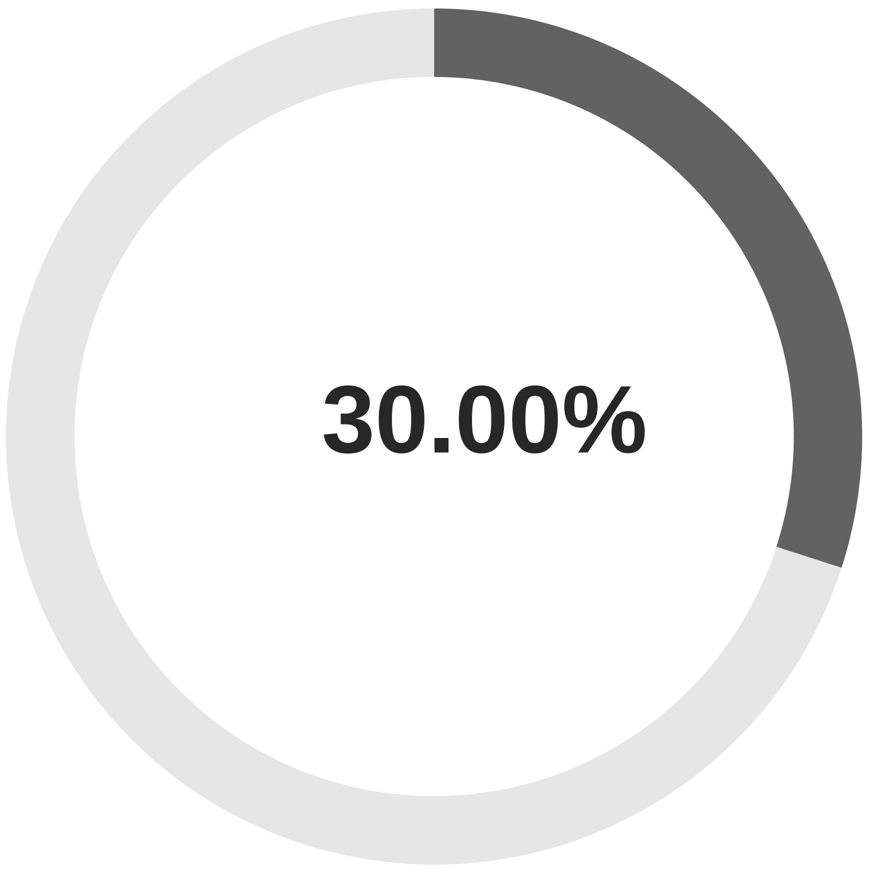
Male



Senior Citizen



Having a Partner



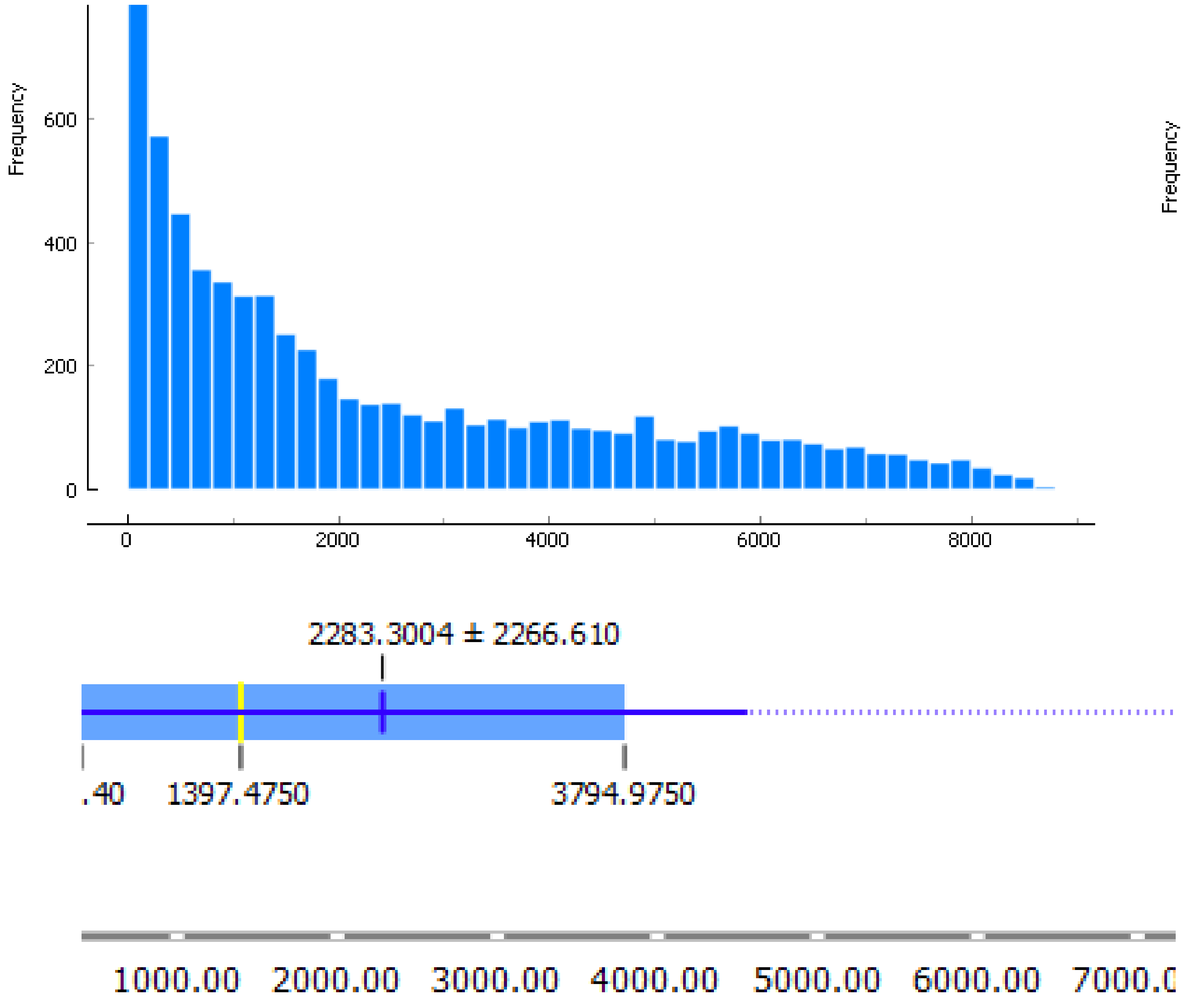
Having a Dependent

Customer Characteristics

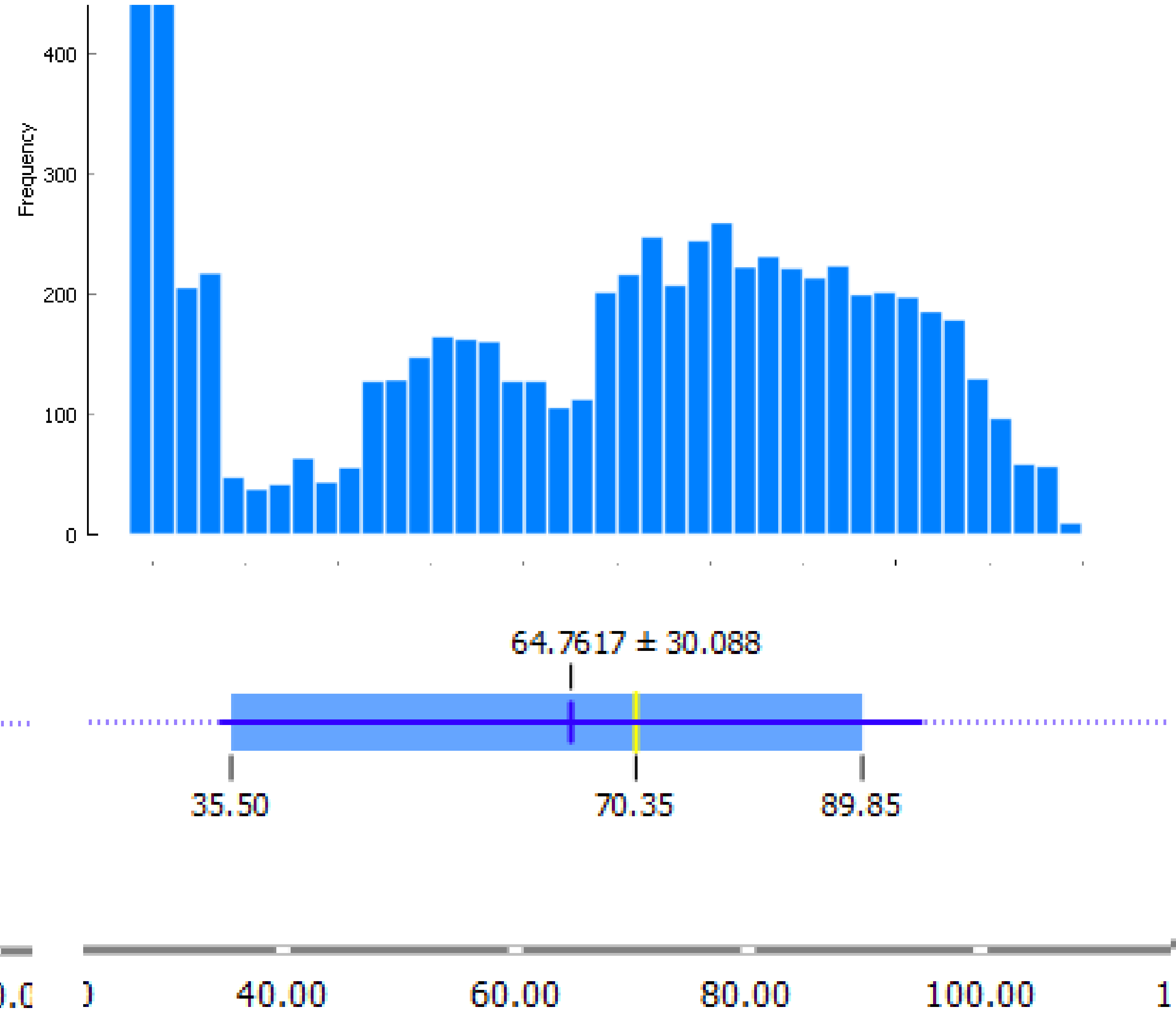
50% of the telco customers,

- **Total charge is less than 1397**
- **Monthly charge is less than 70.35**
- **Stay less than 29 months in network**

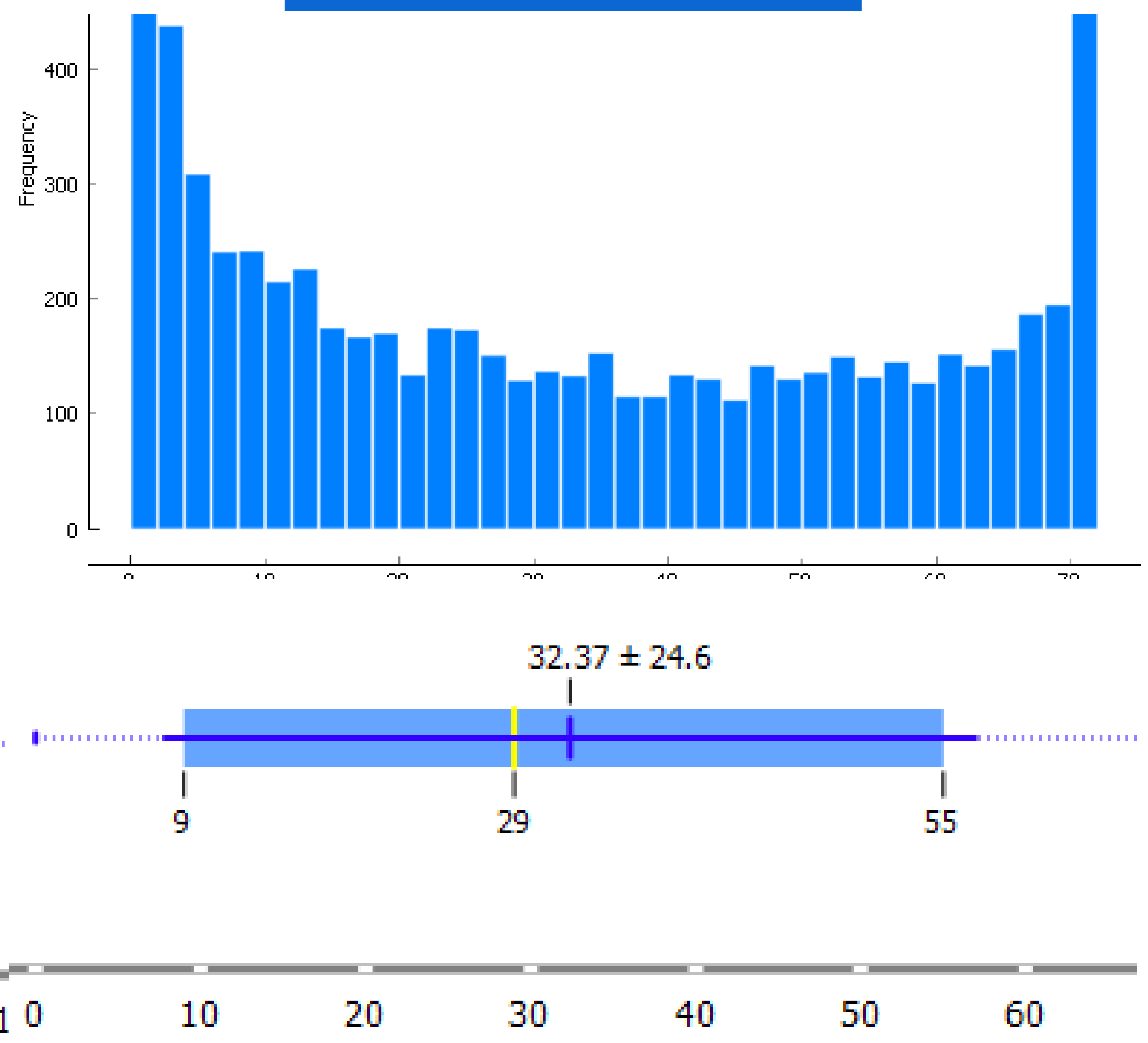
Total Charge



Monthly Charge



Tenure

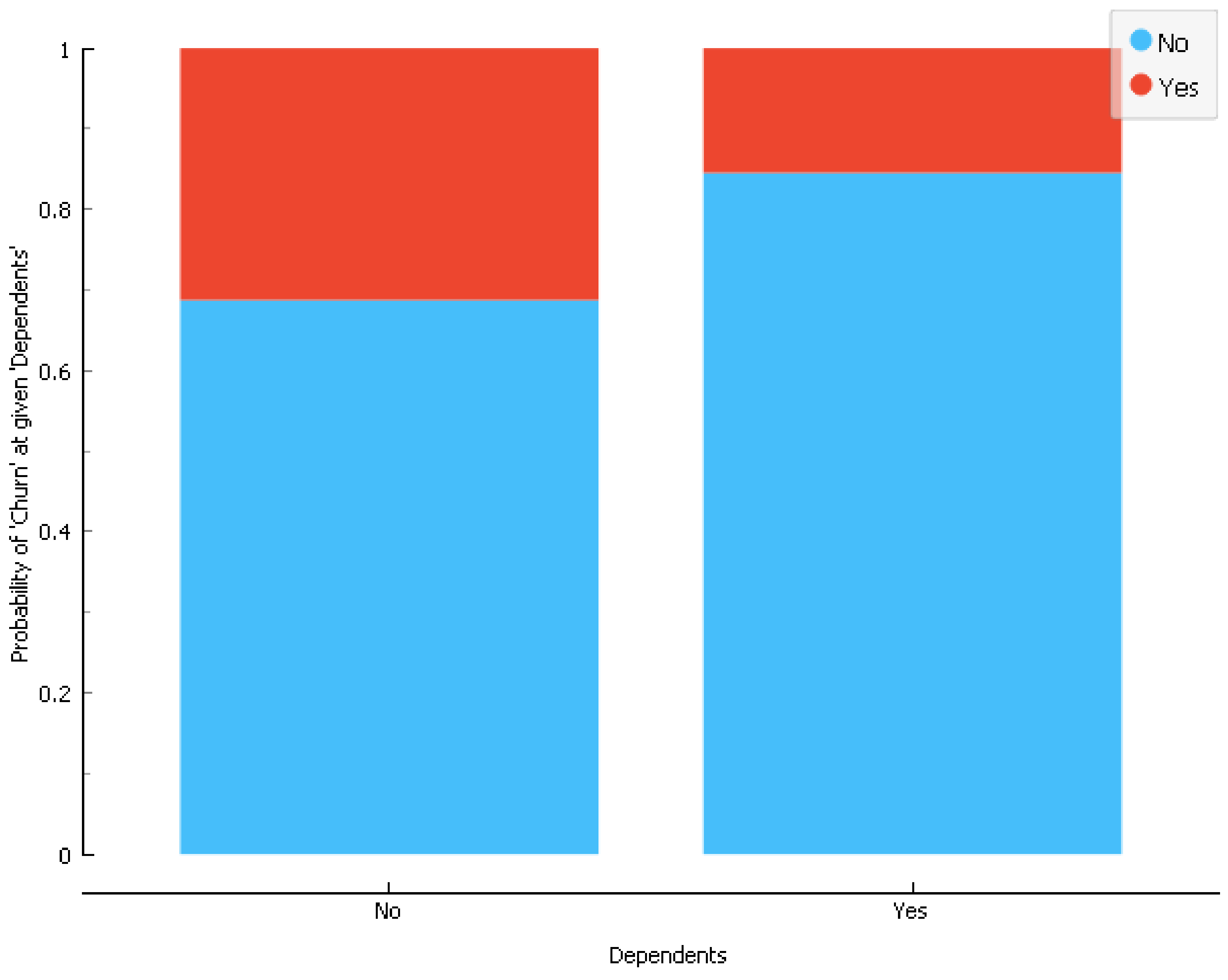


Customers who are more likely to Churn

Demographical Information

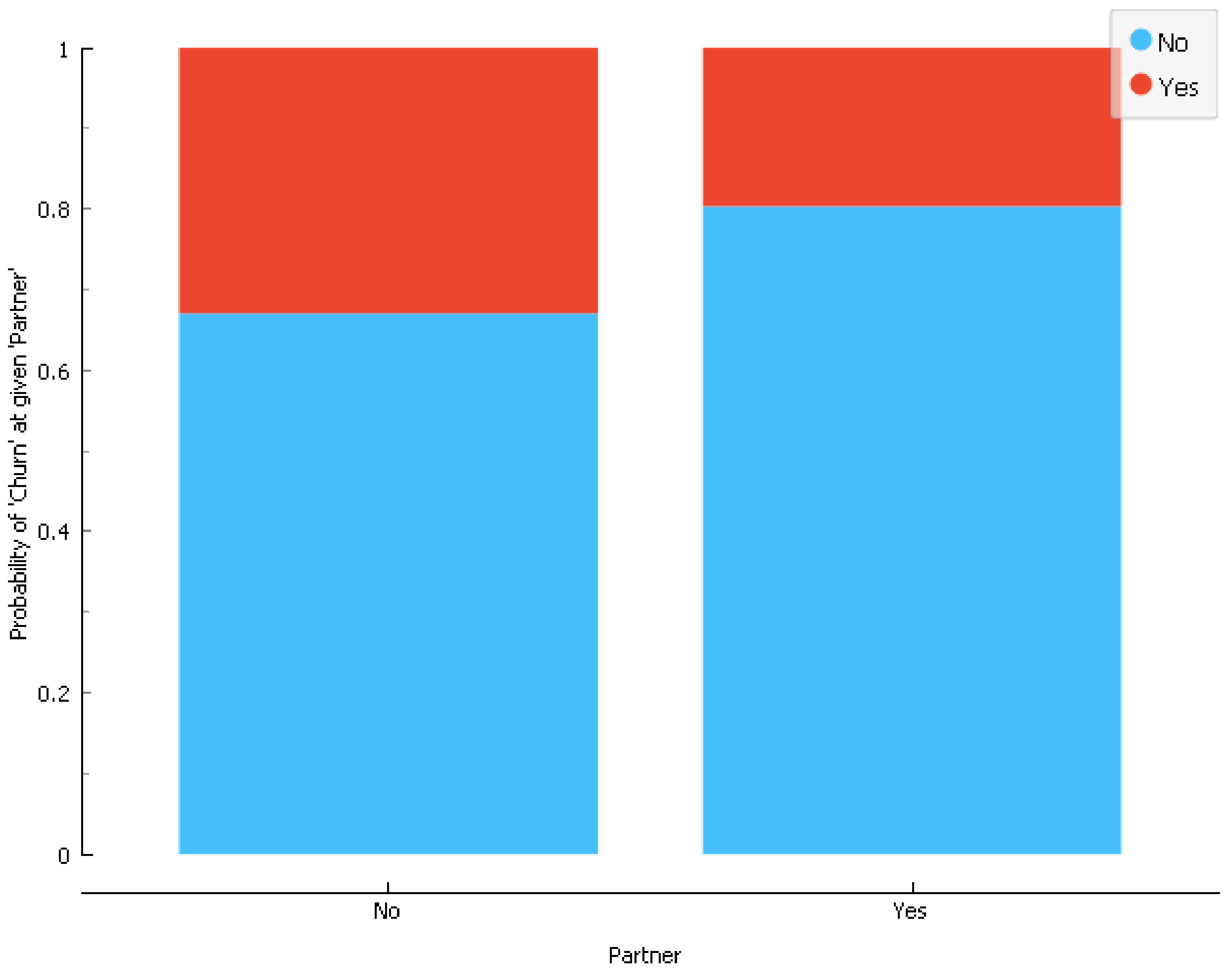
Dependents

Does not Have



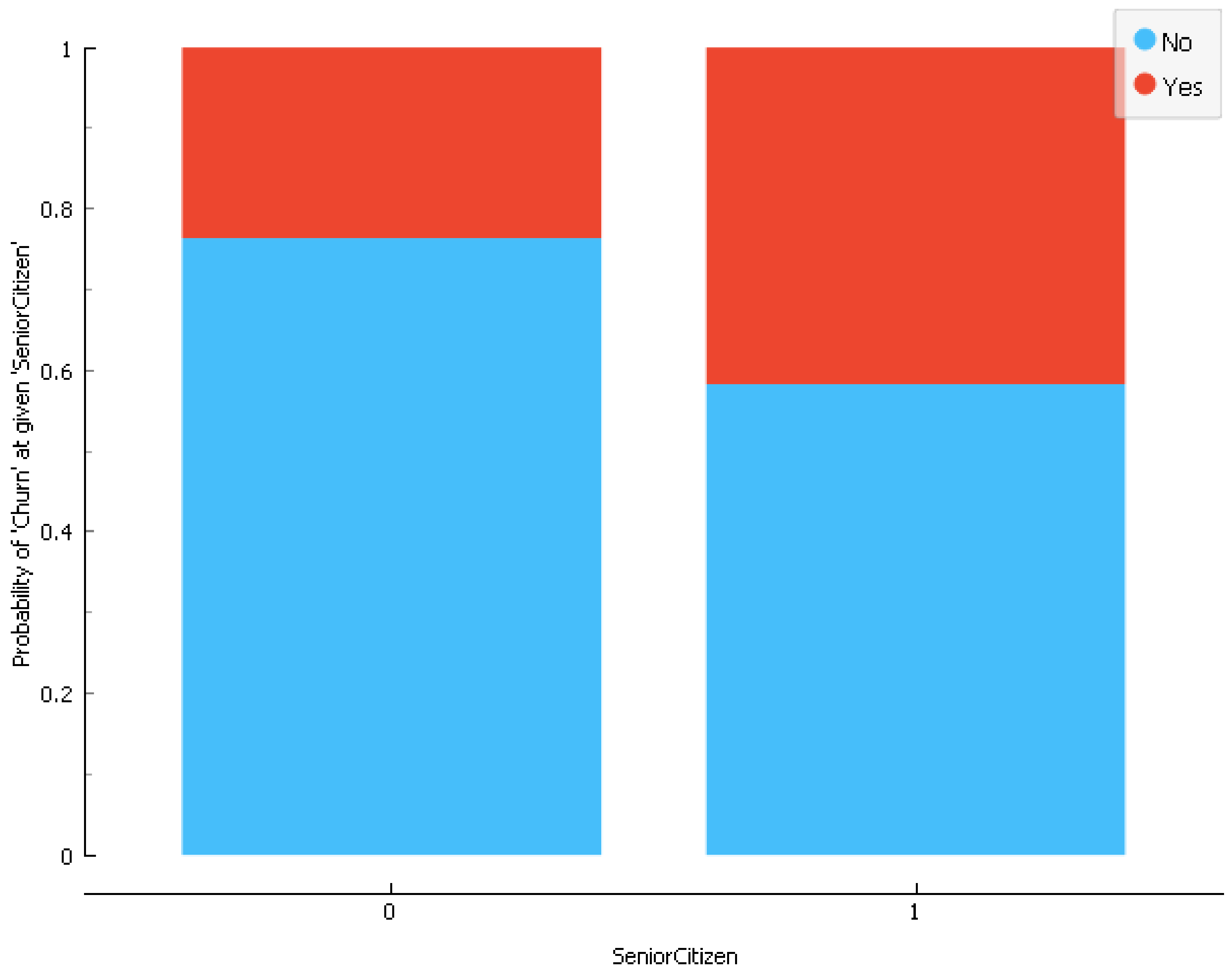
Partner

Does not Have



Senior Citizen

Yes



Customers who are more likely to Churn

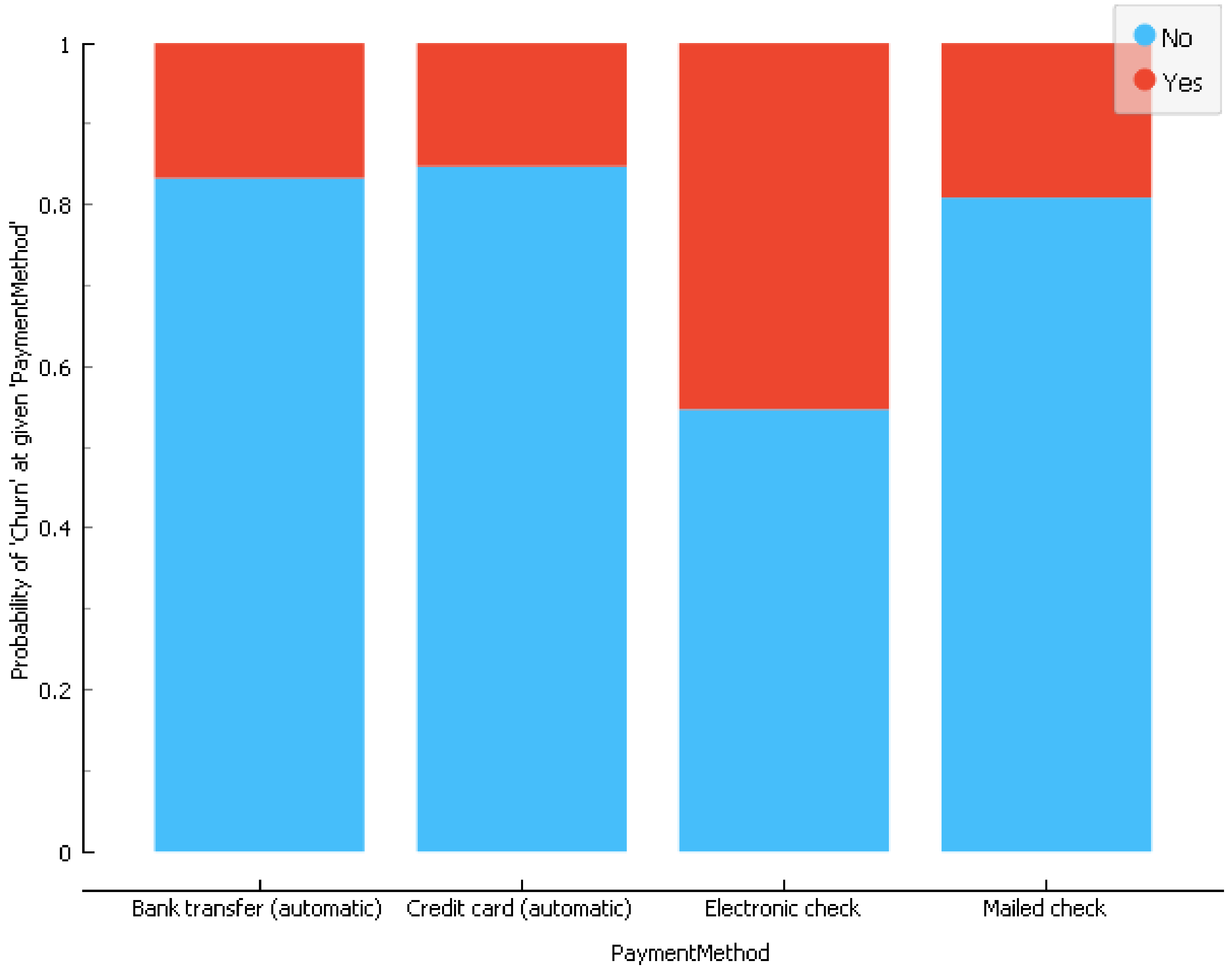
Account Information & Internet Service

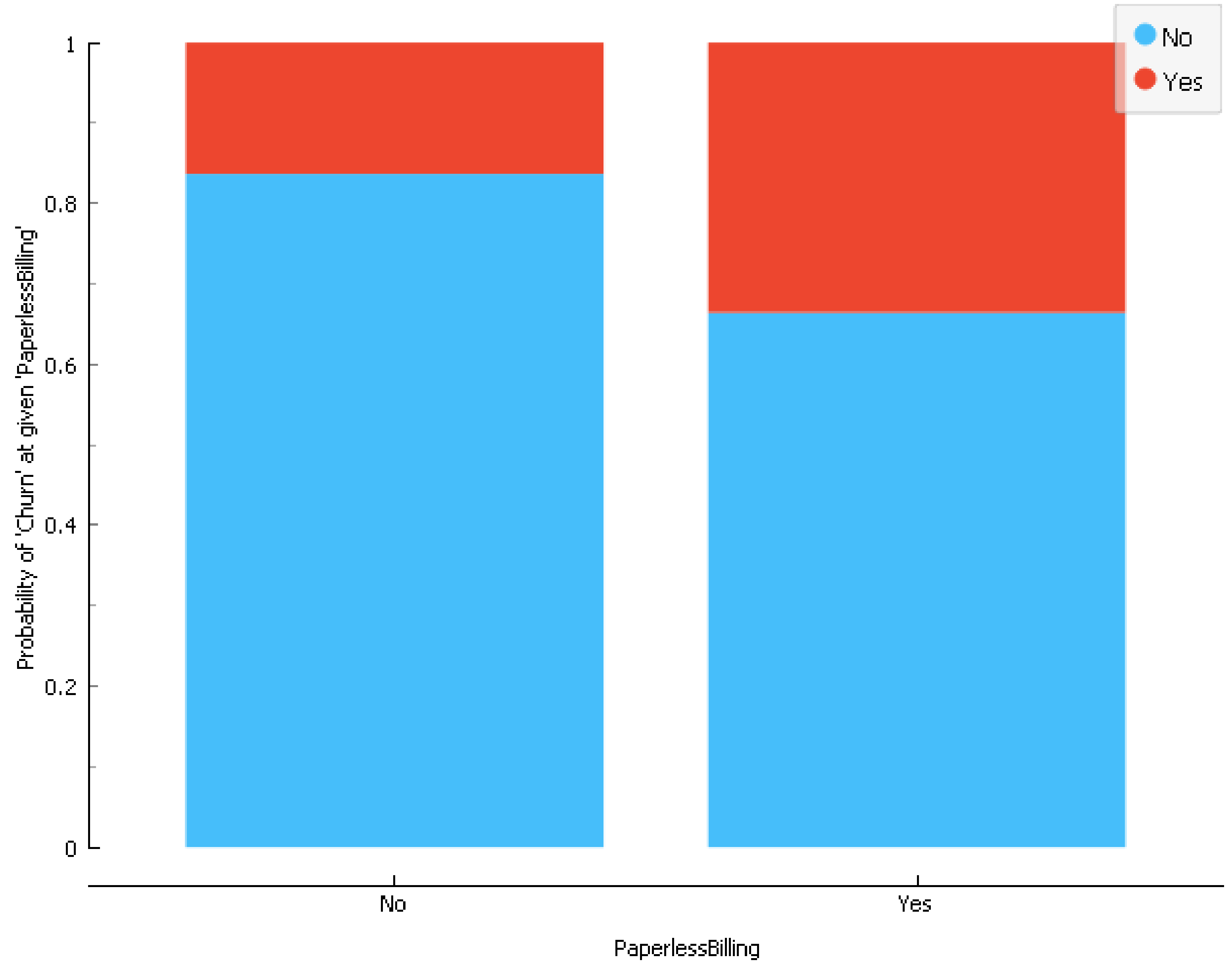
Most of the customers,

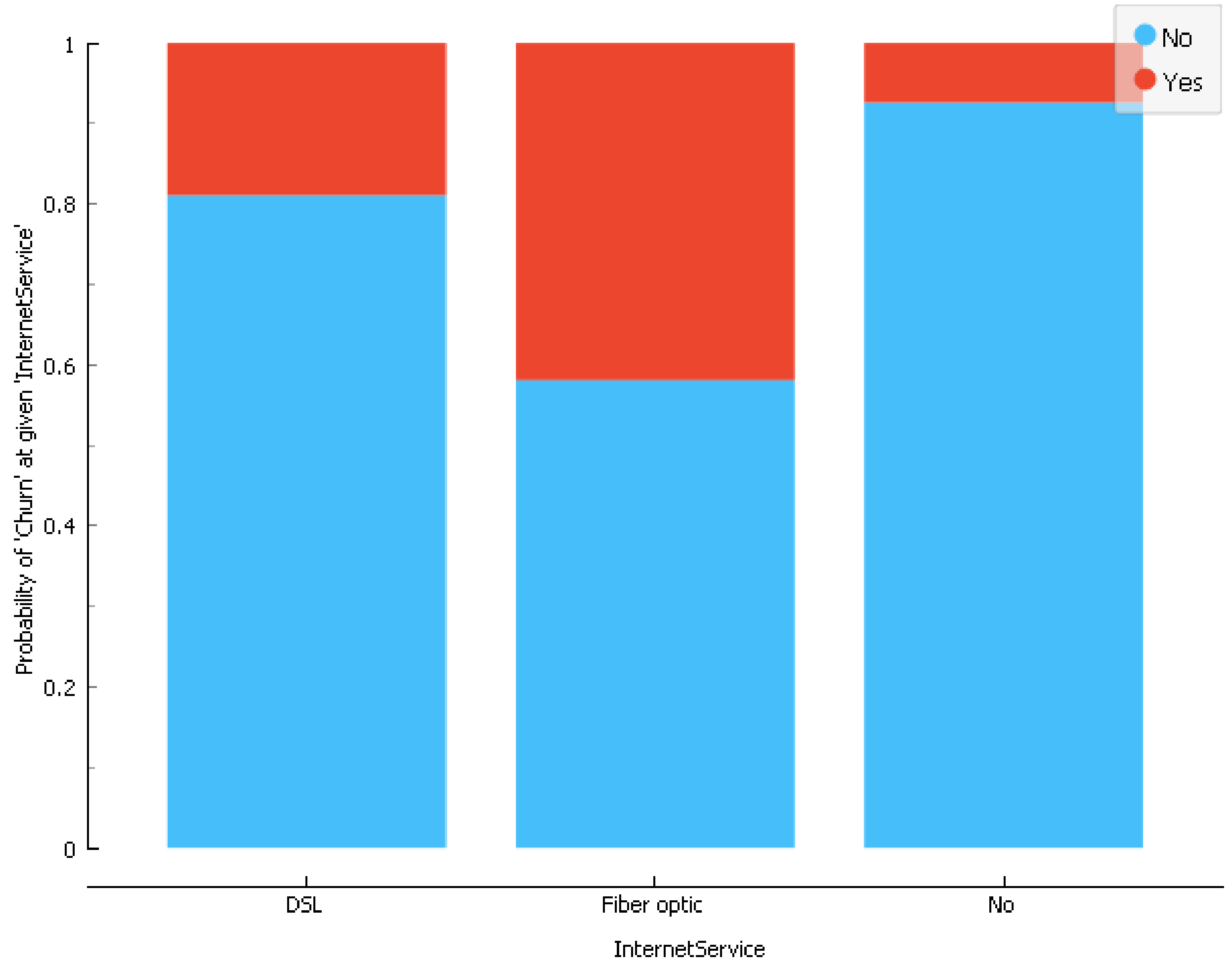
- **use electronic check method as their payment method**
- **subscribed for paperless billing**
- **use fiber optics**

Among these customers, who are more likely to churn,

Payment Method	**Paperless Bill**	**Internet Service**
Electronic Check	Yes (Subscribed)	Fiber Optics

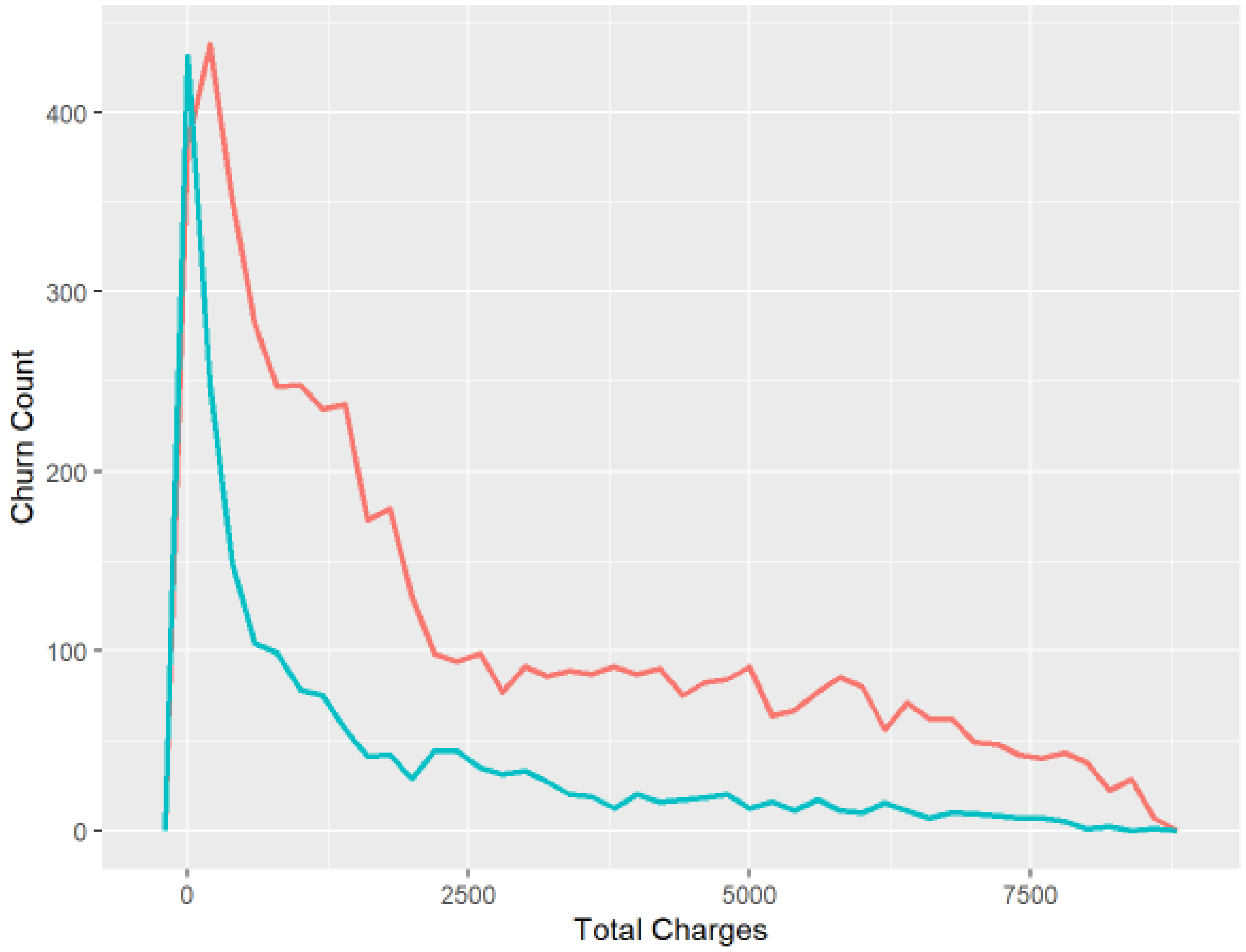






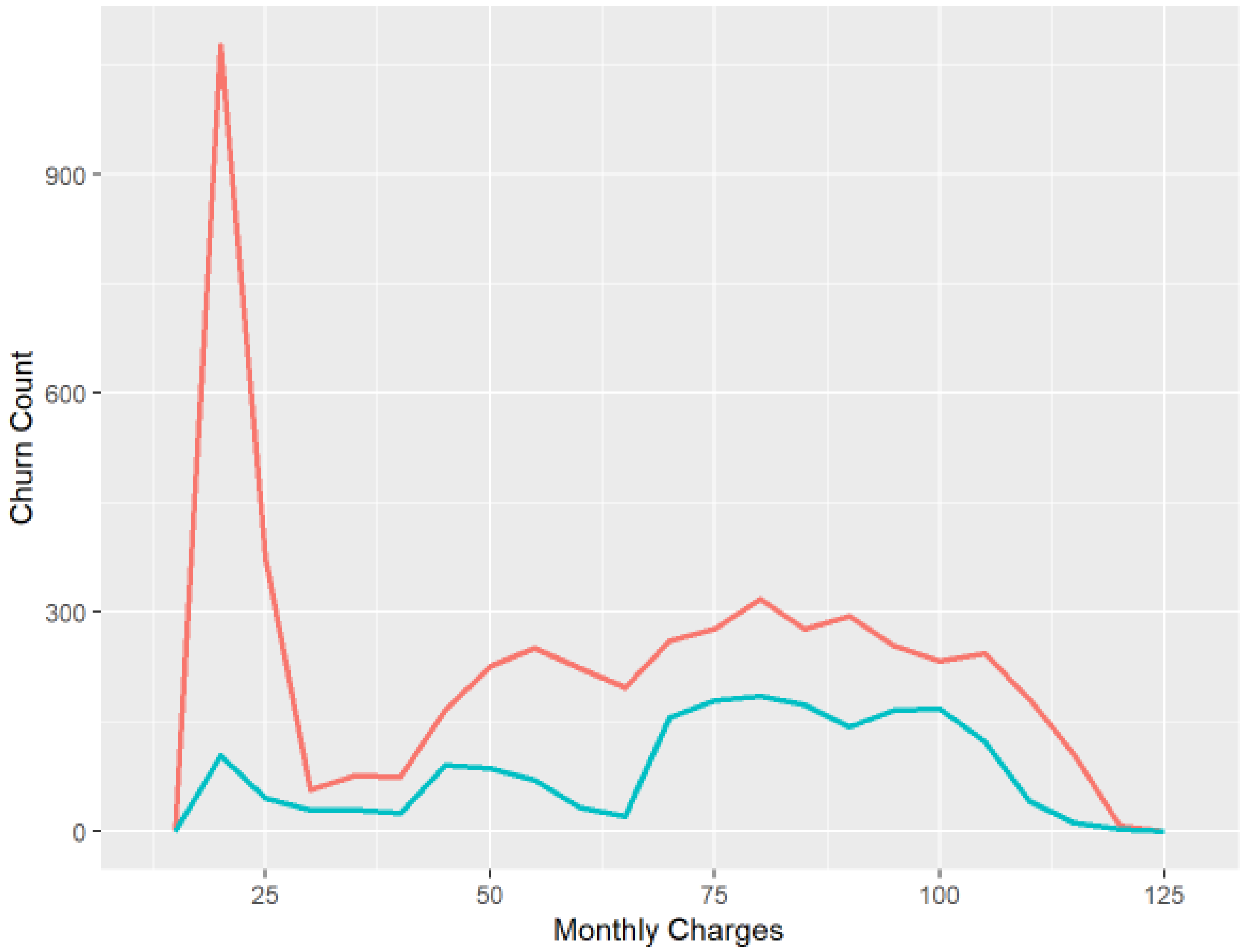
Customers Churn Counts

Churn Vs Total Charge



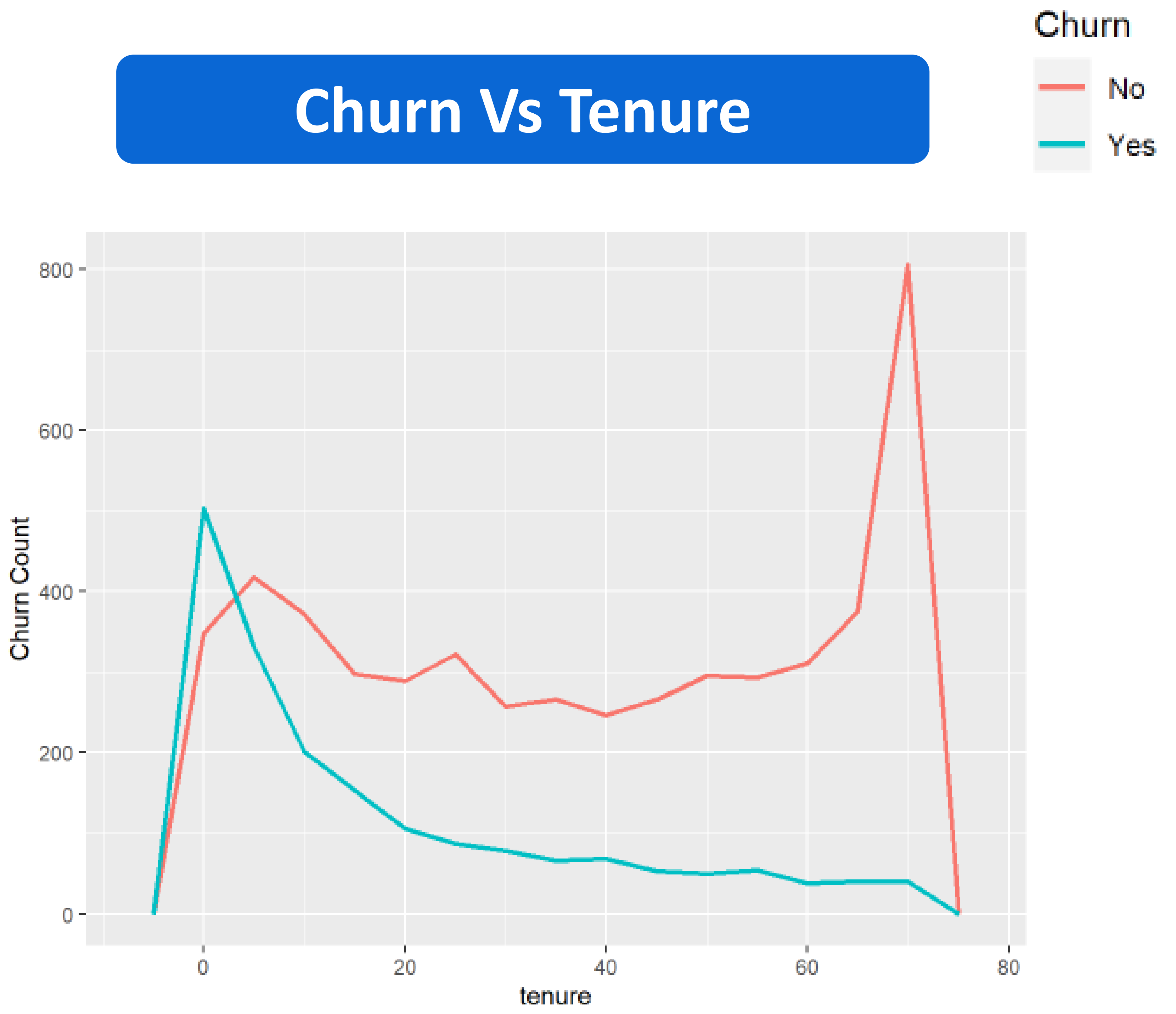
- The total charge two-line charts follow the same pattern for both churned and current customers.
- Although the count of churned customers is low in the dataset, still the count who use below about 250, are same for both categories.

Churn Vs Monthly Charge



- The pattern is different for the customers who have churned from the network.
- The highest frequency of customers belonged to below 25 category.
- More customers who have churned, have used between monthly charges between 75-100.

Churn Vs Tenure

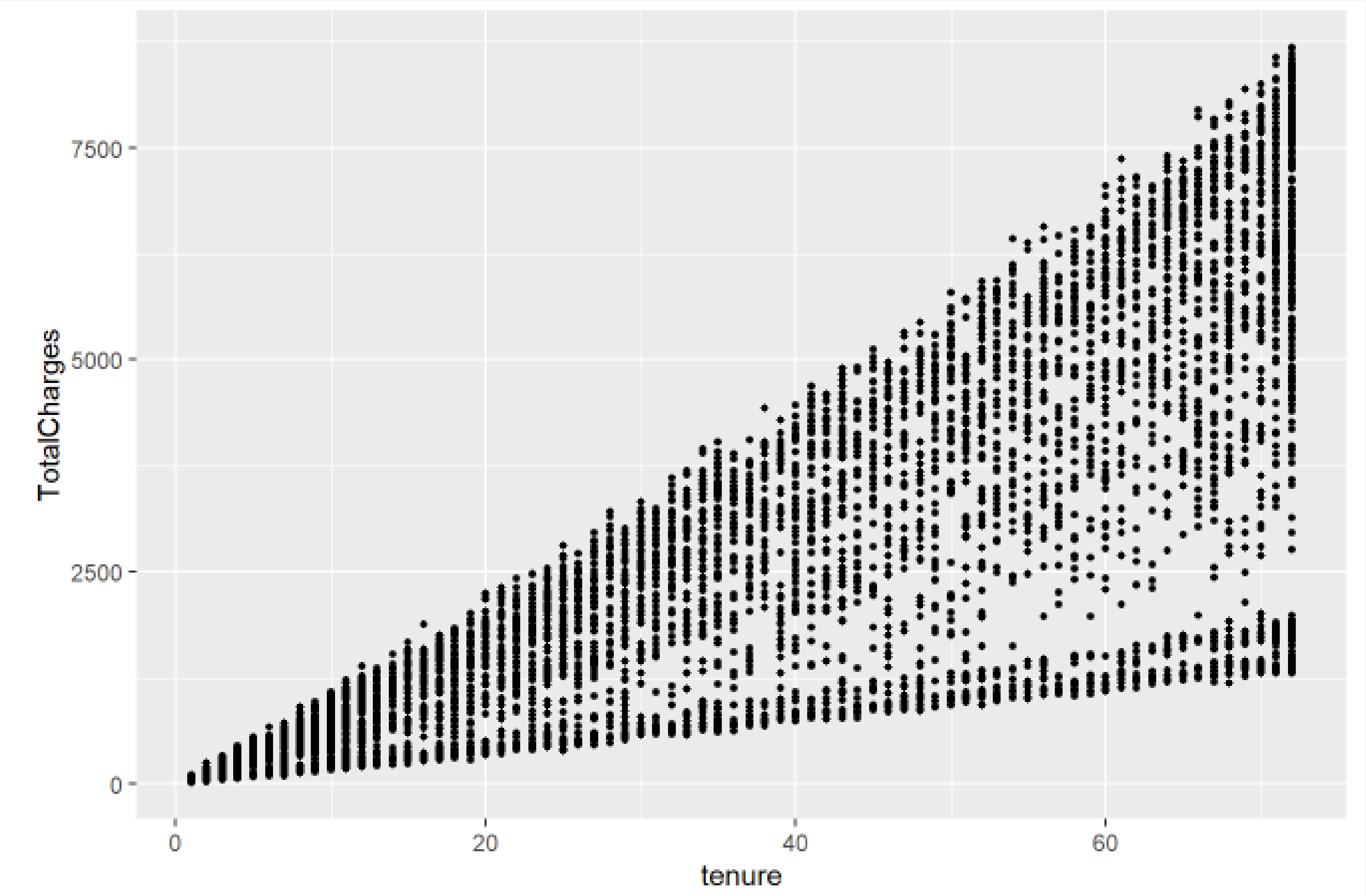


- As the tenure increases count of churn customers are decreasing, while the non churn customers are increasing

Correlation Analysis

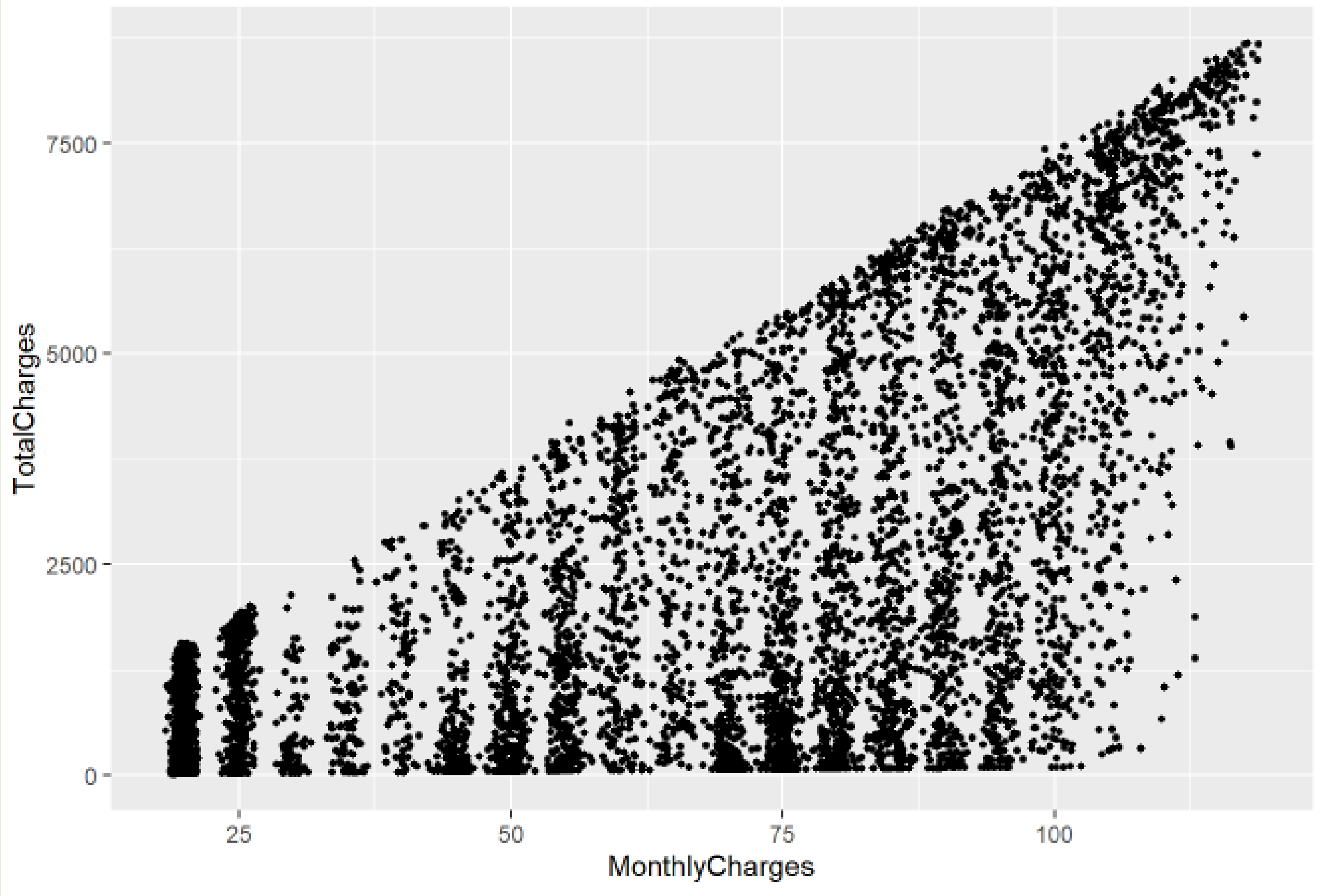
Pearson Correlation Coefficients

Total Charge Vs Tenure



- The correlation coefficient is 0.826.
- There is a linear positive strong relationship between the two factors.

Monthly Charge Vs Tenure



- The correlation coefficient is 0.651.
- There is a linear positive moderate relationship between the two factors.

Correlation Analysis

Overall Correlation Matrix

	tenure	*Monthly Charges*	*Total Charges*
tenure	1.000		
Monthly Charges	0.247	1.000	
Total Charges	0.826	0.651	1.000

- The strongest positive relationship is observed between tenure and total charge (r = 0.826).
- Total charge and Monthly charge have a positive moderate relationship with correlation coefficient of 0651.
- Tenure variable has positive weak relationships with monthly charge (r =0.247).

Data Preprocessing Workflow

We have identified few data quality issues in the data.

- **TotalCarge variable contains 11 missing values.**
- **All the categorical variables in the dataset are in 'char' format**
- **Senior Citizen variable has labelling issue**
- **Data imbalance issue**

We have applied following methods to overcome these issues,

01 Missing Value Replacement

- Only TotalCharges variable contains 11 missing values and it 0.15% of the whole dataset.
- Therefore, we removed those observations from the datasets.

02 Changing the type of the variables

- All the categorical variables in the dataset are in 'char' format.
- Therefore we converted them to factors.

03 Feature Encoding

- Renaming the Senior Citizen variable.
  - 0 = "No
  - 1 = "Yes"

04 Split the dataset

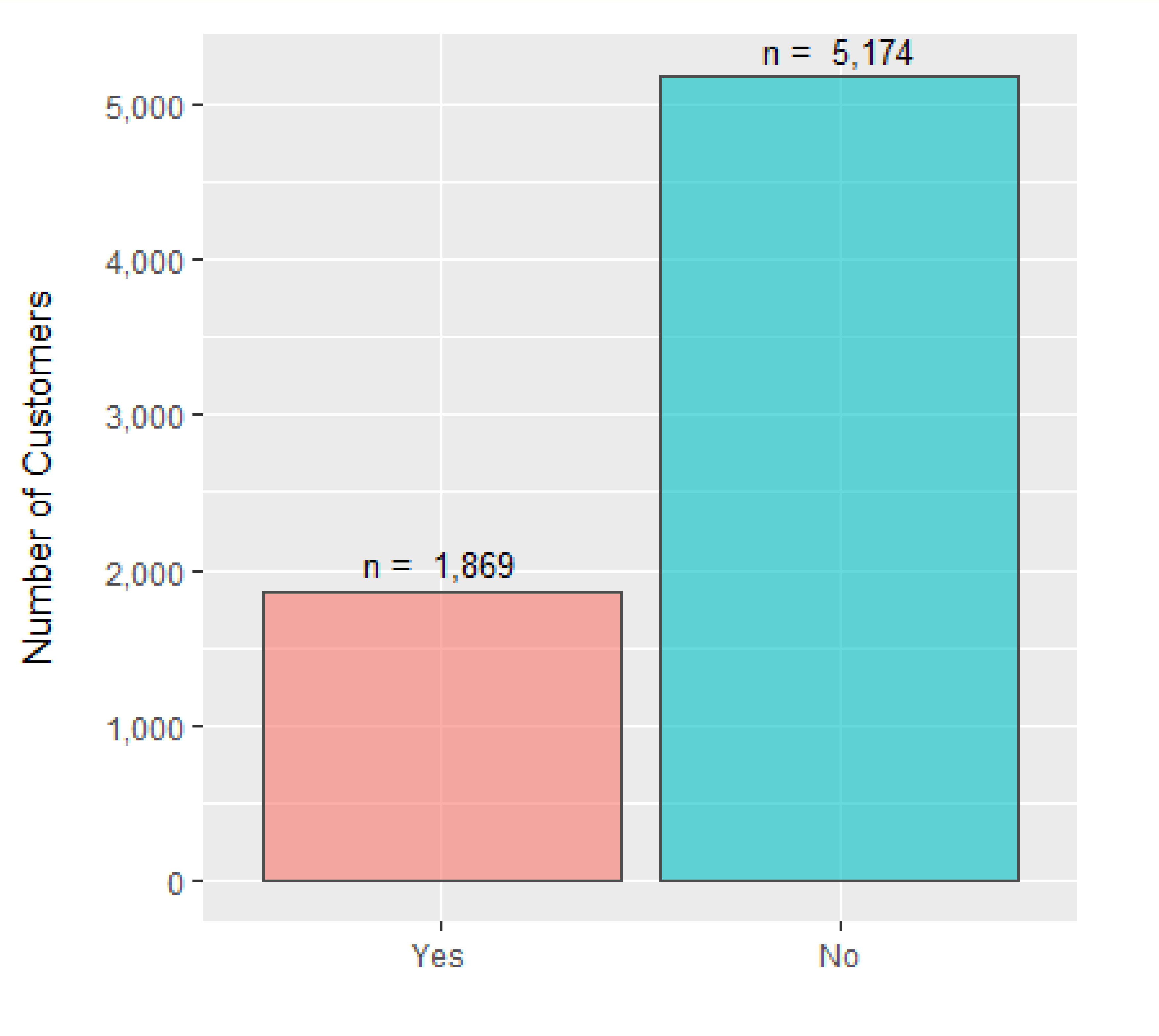
- Split the dataset into training and test sets.
- 70% of the data = training set
- 30% of the data = test set

05 Resampling

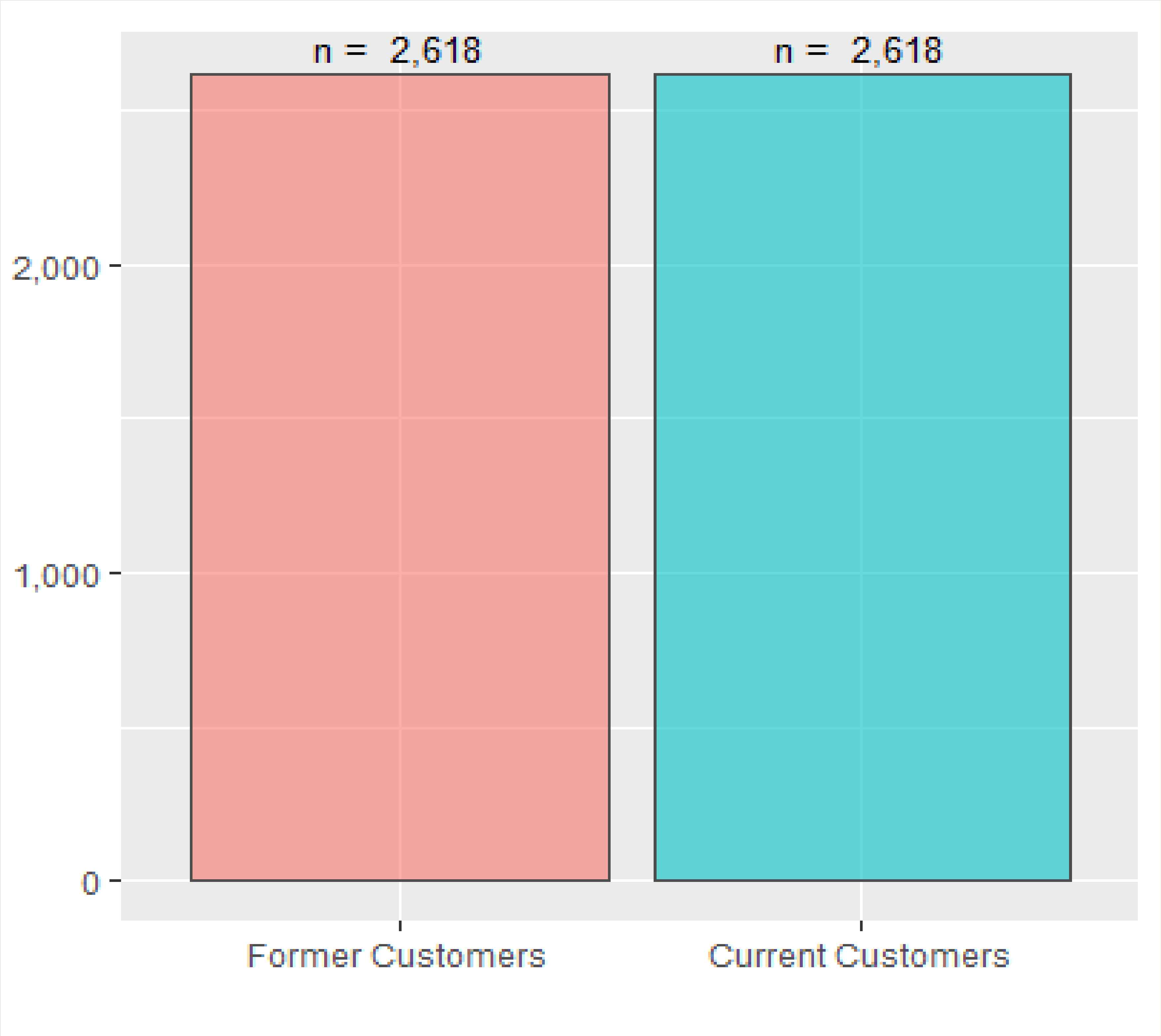
- Churn is quite imbalanced with a little over 26% leaving the company.
- Since class imbalance may negatively affect to the results, we will use SMOTE for training dataset.

Original Dataset vs Resampled Dataset

Original Dataset



After Resampling



Hypothesis Testing : 01

Testing if the average monthly charges differ between males and females

H_0: *The average monthly charge does not differ between males and females* ($\mu_1 = \mu_2$).
H_1: *The average monthly charge does not differ between males and females* (($\mu_1 \neq \mu_2$).

Normality Check

```
 Shapiro-Wilk Normality Test (alpha = 0.05)
--------------------------------------------------
 data : df$MonthlyCharges and gender

  Level Statistic      p.value   Normality
1 Female 0.9215085 2.615068e-39      Reject
2   Male 0.9201174 7.152397e-40      Reject
--------------------------------------------------
```

- The shapiro wilk test also shows a p-value less than 0.001 for both males and females, which indicates the significance of the normality test.

Conclusion: Data are not normally distributed (non parametric)

Wilcoxon rank sum test

```
 Wilcoxon rank sum test with continuity correction

data:  MonthlyCharges by gender
W = 6298265, p-value = 0.249
alternative hypothesis: true location shift is not equal to 0
```

Significance level (=0.05)
P-value = 0.249
As the p-value (=0.249) > 0.05

Fail to reject the null hypothesis.

Conclusion: There is not sufficient evidence to conclude that the average monthly charge for males and females differ.

Hypothesis Testing : 02

Testing if the total charge differs by Payment Method

$$H_0: \mu_1 = \mu_2 = \mu_3.$$
$$H_1: At\ least\ one\ mean\ is\ different$$

Normality Check

```
 Shapiro-Wilk Normality Test (alpha = 0.05)
--------------------------------------------------
 data : df$TotalCharges and PaymentMethod

                     Level Statistic      p.value   Normality
1 Bank transfer (automatic) 0.9228526 2.123728e-27    Reject
2   Credit card (automatic) 0.9175288 5.073903e-28    Reject
3          Electronic check 0.8496637 9.382990e-43    Reject
4              Mailed check 0.7103853 3.209453e-46    Reject
--------------------------------------------------
```

- The Shapiro wilk test shows the p-values for all the categories are less than 0.05 and the tests are significant.

Conclusion: The normality assumption is not met for the data.

Testing for homogeneity

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    3  218.93 < 2.2e-16 ***
      7028
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the Levene's test it is clear that the variances among the four groups are not similar as the p-value is less than 0.001.
The homogeniety assumprion is also not met for this data.

Conclusion: Instead of ANOVA parametric test, Kruskal Wallis non parametric test will be Used.

Hypothesis Testing : 02

Testing if the total charge differs by Payment Method

Applying the Kruskal-Wallis test

```
 Kruskal-Wallis rank sum test

data:  df$TotalCharges by df$PaymentMethod
Kruskal-Wallis chi-squared = 1077, df = 3, p-value < 2.2e-16
```

Significance level (=0.001)
P-value = 2.2e-16
As the p-value < 0.001
Reject the null hypothesis.

Conclusion: at least one mean total charge is different from others..

Posthoc Analysis

```
 Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data:  df$TotalCharges and df$PaymentMethod

                        Bank transfer (automatic) Credit card (automatic)
Credit card (automatic) 0.7                       -
Electronic check        <2e-16                    <2e-16
Mailed check            <2e-16                    <2e-16
                        Electronic check
Credit card (automatic) -
Electronic check        -
Mailed check            <2e-16

P value adjustment method: BH
```

Conclusion: The post hoc analysis shows that except between bank transfers and credit card payments, all the mean total charges are different for all categories.

Hypothesis Testing : 03

Testing if the payment method and the contract type are independent

H_0: *payment method and the contract type are independent.*
H_1: *payment method and the contract type are not independent.*

Applying Chi-Square Test

```
 Pearson's Chi-squared test

data:  obs
X-squared = 1001.6, df = 6, p-value < 2.2e-16
```

- The p-value obtained from the chi squared test for independence is less than 0.001.
- Therefore the null hypothesis is rejected.

Assumption Checking

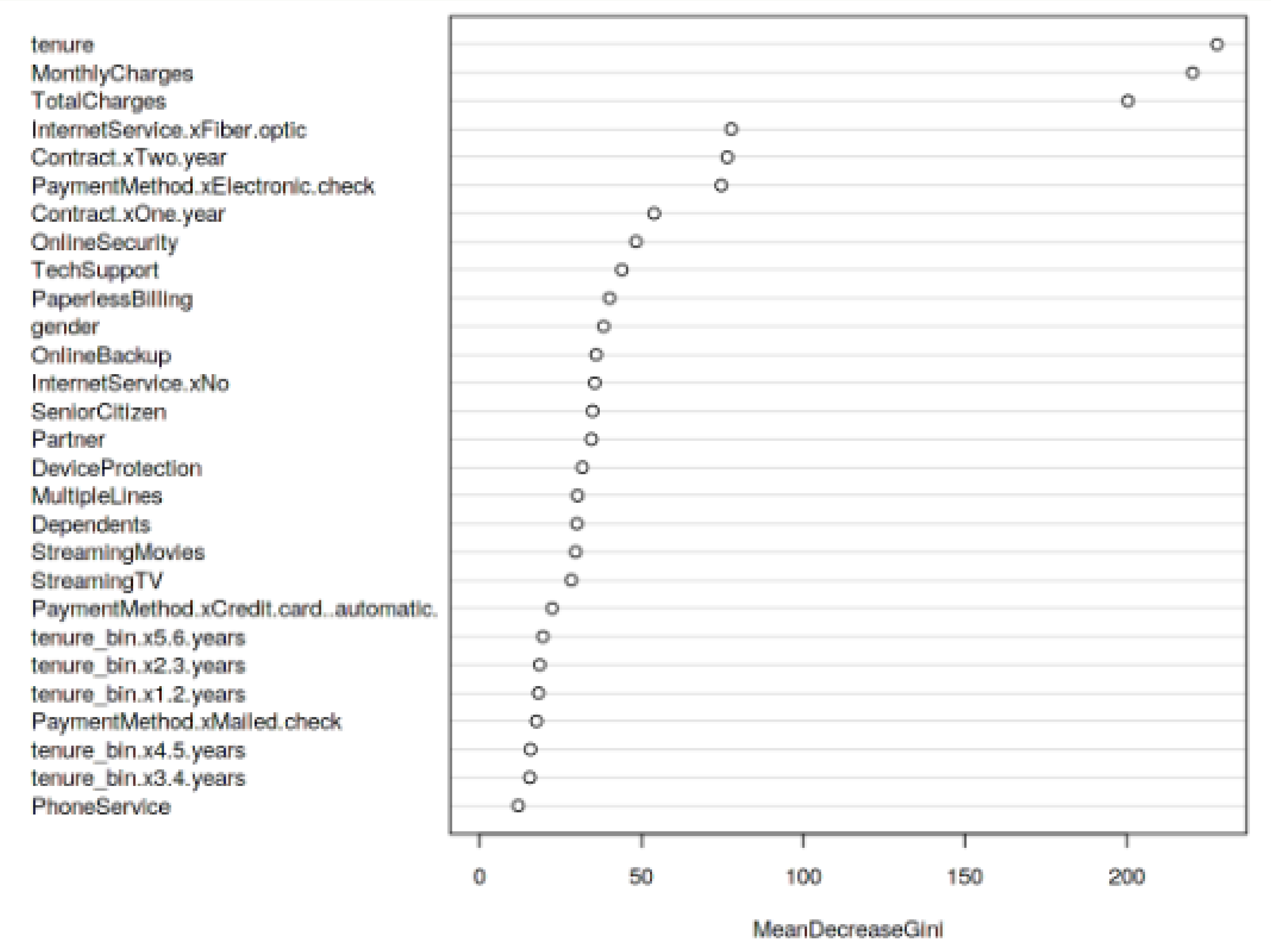
	Month-to-month	One year	Two year
Bank transfer (automatic)	849.4960	322.9181	371.5860
Credit card (automatic)	837.3917	318.3169	366.2914
Electronic check	1301.2033	494.6252	569.1715
Mailed check	886.9090	337.1399	387.9512

The expected values are greater than 5 for each cell. Therefore, the expected value at least be 5, assumption is met for the data.

Conclusion: There is sufficient evidence to conclude that the payment method and the contract type are independent.

Feature Selection

Feature Importance



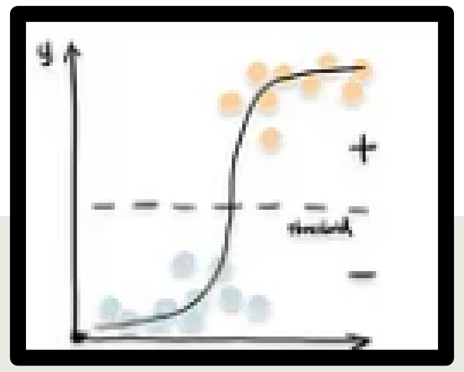
- Most important variables are Tenure, MonthlyCharge and TotalCharge.
- Due to significant correlation (multicollinearity) between tenure and totalcharge, only selected tenure for the analysis.

- All the features used for the analysis except **totalCharge** variable.

Model Training

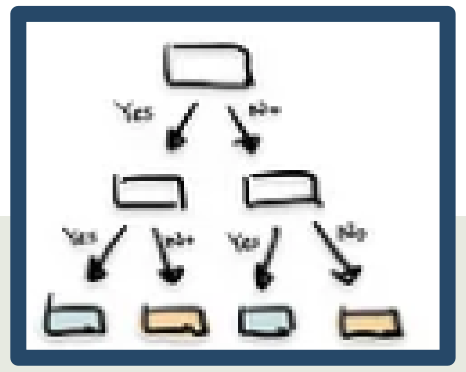
The binary classification technique was used for the resampled dataset.

Following models are applied to the resampled dataset.



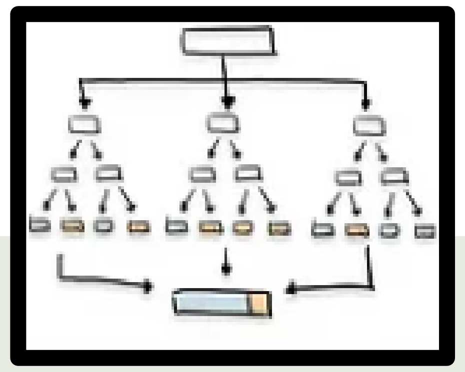
Logistic regression is a process of modeling the probability of a discrete outcome given an input variable.

Logistic Regression



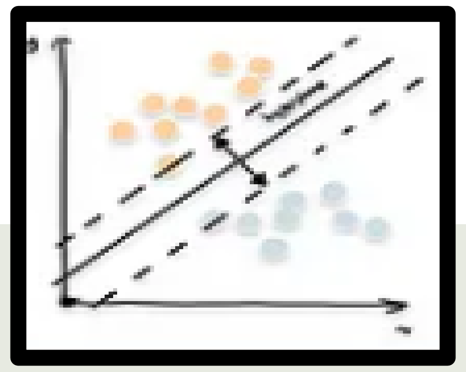
A decision tree is a graph that uses a branching method to illustrate every possible output for a specific input.

Decision Tree



The random forest is a supervised learning algorithm that randomly creates and merges multiple decision trees into one "forest."

Random Forest



Support vector machines (SVMs) are supervised learning models that analyze data and recognize patterns, used for classification and regression analysis

Support Vector Machines (SVM)

Model Evaluation

- Performances of each model are mainly measured using **Accuracy, Precision, Recall and F1-score**.
- But the performance can not directly be assessed by using the accuracy measure as the imbalances of the dataset.
- Hence, the F1-score, recall and precision are the most appropriate measurement to evaluate the model.
- Among these measures, high priority is given to **the recall and f1 score**. Therefore the process is recall-oriented.

	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.7663	0.5358	0.8143	0.6464
Decision Tree	0.7381	0.8375	0.5043	0.6295
Random Forest	0.7533	0.5238	0.7857	0.6286
Support Vector Machines (SVM)	0.6997	0.4644	0.8500	0.6006

Best model

Best Model

- Logistic regression has the best performance on the test set and below summary shows, the coefficients and the performances on the test set.

Logistic Regression Model

$\widehat{churn}$ = *-0.13573 -0.78385 tenure + 0.75276 MonthlyCharges + 0.19104 InternetserviceFiberoptic +0.03143 InternetServiceno -0.06496 PaymentMethodcreditcardauto + 0.21314 PaymentMethodecheck – 0.03878 PaymentMethodMailedcheck - 0.31086 ContractoneYear - 0.56591 ContractTwoYear - 0.16972 onlinesecurityyes - 0.21416 TechsupportYes + 0.14521 Paperlesssillingres – 0.04852 genderMale -0.26178 PhoneServiceYes*

How will you gauge if this model is better able to predict customer churn ?

Model Performance

Accuracy	0.7633	Recall	0.8143
Precision	0.5358	F1 Score	0.6464

Confusion matrix

		Reference	
Prediction		No	Yes
	No	1148	120
	Yes	400	440

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -0.13573    0.03761  -3.609 0.000308 ***
tenure                      -0.78385    0.05794 -13.529  < 2e-16 ***
MonthlyCharges               0.75276    0.11233   6.701 2.06e-11 ***
InternetServiceFiberOptic    0.19104    0.07252   2.634 0.008430 **
InternetServiceNo            0.03143    0.06762   0.465 0.642105
PaymentMethodCreditCardAuto -0.06496    0.04599  -1.413 0.157768
PaymentMethodECheck          0.21314    0.05009   4.255 2.09e-05 ***
PaymentMethodMailedCheck    -0.03878    0.04749  -0.817 0.414191
ContractOneYear             -0.31086    0.04092  -7.596 3.06e-14 ***
ContractTwoYear             -0.56591    0.06198  -9.131  < 2e-16 ***
OnlineSecurityYes           -0.16972    0.03894  -4.359 1.31e-05 ***
TechSupportYes              -0.21416    0.04044  -5.296 1.18e-07 ***
PaperlessBillingYes          0.14251    0.03796   3.755 0.000174 ***
genderMale                  -0.04852    0.03479  -1.394 0.163179
PhoneServiceYes             -0.26178    0.04438  -5.899 3.67e-09 ***
```

Best Model

Logistic Regression Model

$\widehat{churn}$ = -0.13573 -0.78385 tenure + 0.75276 MonthlyCharges + 0.19104 InternetserviceFiberoptic +0.03143 InternetServiceno -0.06496 PaymentMethodcreditcardauto + 0.21314 PaymentMethodecheck – 0.03878 PaymentMethodMailedcheck - 0.31086 ContractoneYear - 0.56591 ContractTwoYear - 0.16972 onlinesecurityves - 0.21416 TechsupportYes + 0.14521 Paperlesssillingres – 0.04852 genderMale -0.26178 PhoneServiceYes

Interpret the coefficients

- -0.78385 tenure

  Every one month increase in tenure, expected churn decreases by 0.78385

- +0.75276 MonthlyCharge

  Every one dollar increase in monthly charge, expected churn increase by 0.75276

Importance of the Best model

The model can be used to:

- Identify customers who are likely to churn
- Can develop this model to identify the high-risk customer
- Segmenting the customers based on the churn propensity (high risk/medium risk/low risk)

- Identify the most important features of customer churn
- Based on the results company can offer special promotions campaigns to high risk/ identified potential churn customers
- Can include this predictions base as an input variable for other ongoing machine learning projects

Limitations of our Best model

May not be able to capture complex non-linear relationships between variables

Assumes that the relationships between the independent and dependent variables are always linear

May not be accurate for a huge amount of data

Assume variables followed normal distribution and independent

The model can be prone to overfitting

Our model may not be well-suited for predicting customer churn in cases where the data is time-series in nature

Limitations

The limitations of the dataset:

- No of observations in the dataset is relatively small.
- Lack of variables in the dataset
- Do not have time series factors/features in the dataset.
- Imbalanced data for some classes
- Outdated data

Suggestions

Suggestions on how the dataset can be improved:

- Increasing the sample size by collecting more data
- Adding more features to the dataset like customer’s network usages, competitor’s details, etc.)
- Adding time series data to the dataset
- Balancing the class distribution
- Regularly collecting and updating the data

Conclusion

- Dataset consists 7043 observations and among those observations, most of them are male, that are not senior citizens and do not have a dependent or partner.
- Telco company should target the customer who is a senior citizen, does not have a partner or dependent, use fiber optics and payment done by electronic check
- **Tenure, MonthlyCharge, Contract, PayementMethod** and **OnlineSecurity** are the top 5 most important features of the analysis,
- Dataset is imbalanced for some classes and there is a strong positive relationship between **tenure** and **TotalCharge** variables (multicollinearity occurred)
- Logistic regression model gave the best performances among the other algorithms


Q
&
A
?