Why My Customers Are Leaving?

Muhammad Fawzi Al-Akhdar

2025-02-23

Contents

Abstract

Imagine a telecom provider losing 5% of its customers monthly, one way is to research examine the factors that lead to customer churn in the telecommunications industry, we did that with a case study of a telecommunications company. Using exploratory data analysis and predictive modeling, key variables such as SMS usage, calling plan, and customer age group are examined for their impact on churn. Logistic regression modeling with cross-validation provides insight into the accuracy of churn prediction and guides targeted retention strategies. Limitations and future enhancements to improve model accuracy and practical application are discussed.

Introduction

Customer attrition, or churn, is a critical challenge for businesses, especially in competitive industries like telecommunications.**"Becuase retaining existing customers is more profitable than acquiring new customers, primarily due to savings in acquisition costs, higher volume of service consumption, and customer referrals."**

For a telecom company based in Iran, building an effective customer retention program can reduce churn.To do so they can use their **Dataset**, to uncover patterns to understand why customers leave and identify those at high risk of leaving by accurately predicting customer churn so they can target them. By carefully analyzing and digging deeper into the dataset, we can predict and understand customer churn.

Methodology

Data Cleaning

Consistency of the dataset is important for meaningful data analysis, and we will ensure it.

Our steps include: 1. Rename columns for readability. 2. Convert binary and categorical variables to factors. 3. Check for missing values for data completeness.

Here is a snapshot of the original dataset:

Table 1: A snapshot of the orginal dataset

Call..Failure	Complains	Subscription..Length	Charge..Amount	Seconds.of.Use	Frequency.of.use
11	0	40	1	6223	101
9	0	38	1	4213	89
15	0	18	1	5987	136
0	0	37	0	5208	84
0	0	26	0	825	10

Here is a snapshot of the dataset after cleaning:

Table 2: A snapshot of the cleanded dataset

Call Failure	Complains	Subscription Length	Charge Amount	Seconds Of Use	Frequency Of Use
13	No complaint	40	1	4313	71
5	No complaint	30	2	6885	114
1	No complaint	36	0	10755	143
0	No complaint	31	0	5880	81
18	No complaint	38	2	9305	121

Table 3: A snapshot of the cleanded dataset

Frequency Of Sms	Distinct Called Numbers	Age Group	Tariff Plan	Status	Age
0	34	1	Pay as you go	Acitve	15
261	29	2	Contractual	Acitve	25
0	47	3	Pay as you go	Acitve	30
0	5	4	Pay as you go	Acitve	45
119	41	3	Pay as you go	Acitve	30

Table 4: A snapshot of the cleanded dataset

Customer Value	Churn

241.120	non-churn
1489.455	non-churn
435.920	non-churn
149.025	non-churn
853.040	non-churn

Exploratory data analysis (Why Customers Leave?)

- Identifying a criteria that distinguishes customers is our goal in this section.

Although we will not explore every avenue of the dataset, we will start by asking a few questions and then refine our questions as we dig deeper:

1. What are the most and least common values within the data?
2. Are there any unusual patterns?
3. Are there any correlations within the data? Will start with calculated customer value in customers who left.

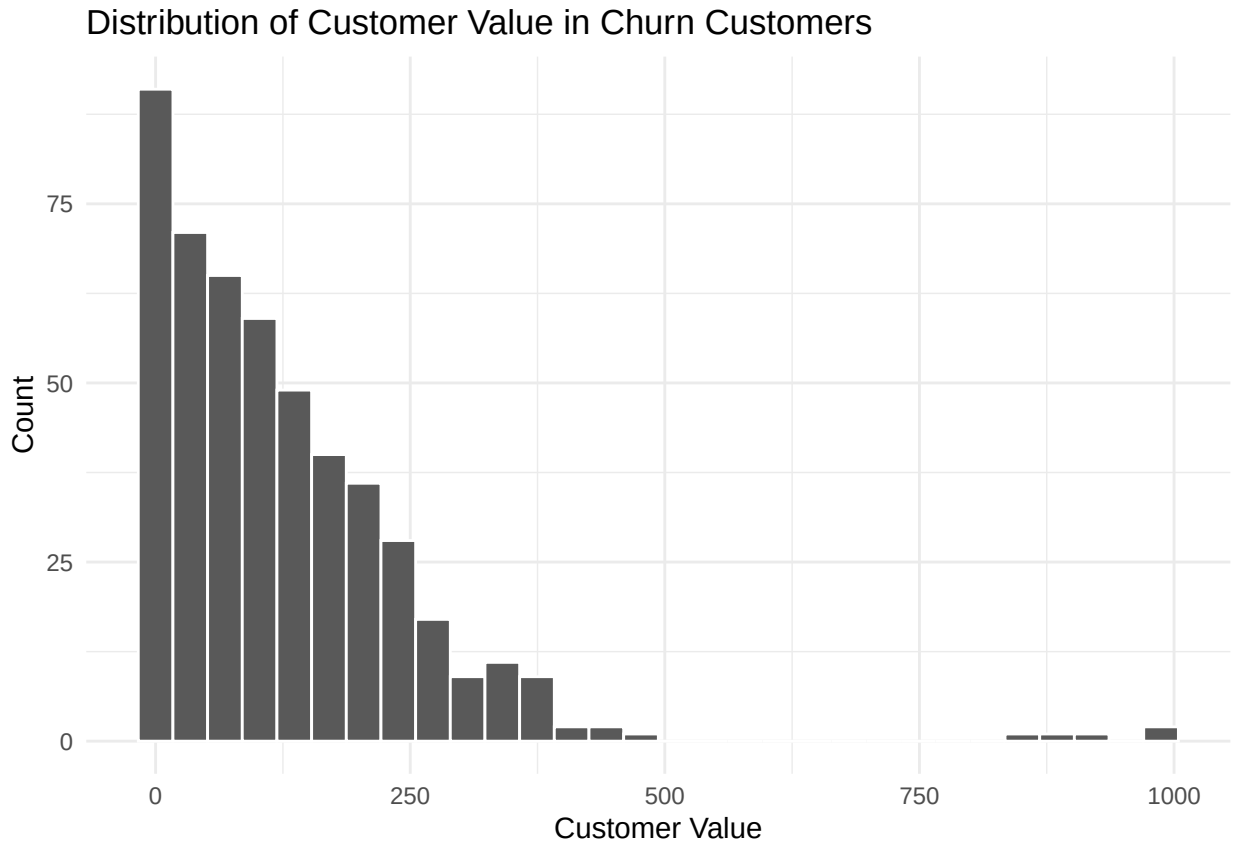
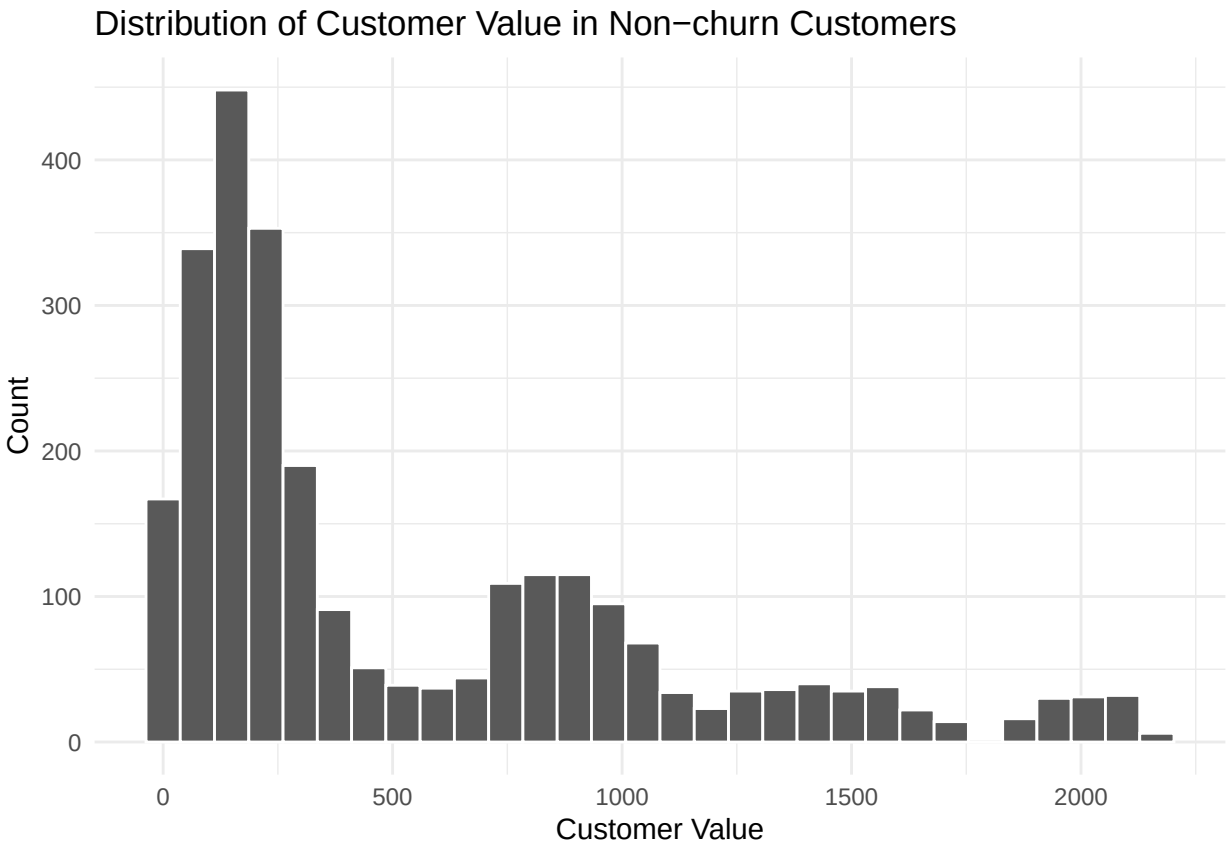


Figure 1: Distribution of Calculated Customer Value and Churn

- Figure 1 shows that the most common values is zero which is not expected here it could be due to a missing value or real effect. Something to notice here that the figure is skewed to the right which in return might not represent the customers as they are we will address this problem later.

- Next we will see the most common values in non-churn customers.



- Figure 2 shows clustering or subgroups of customers with same values the most common value seems to be less than 500.

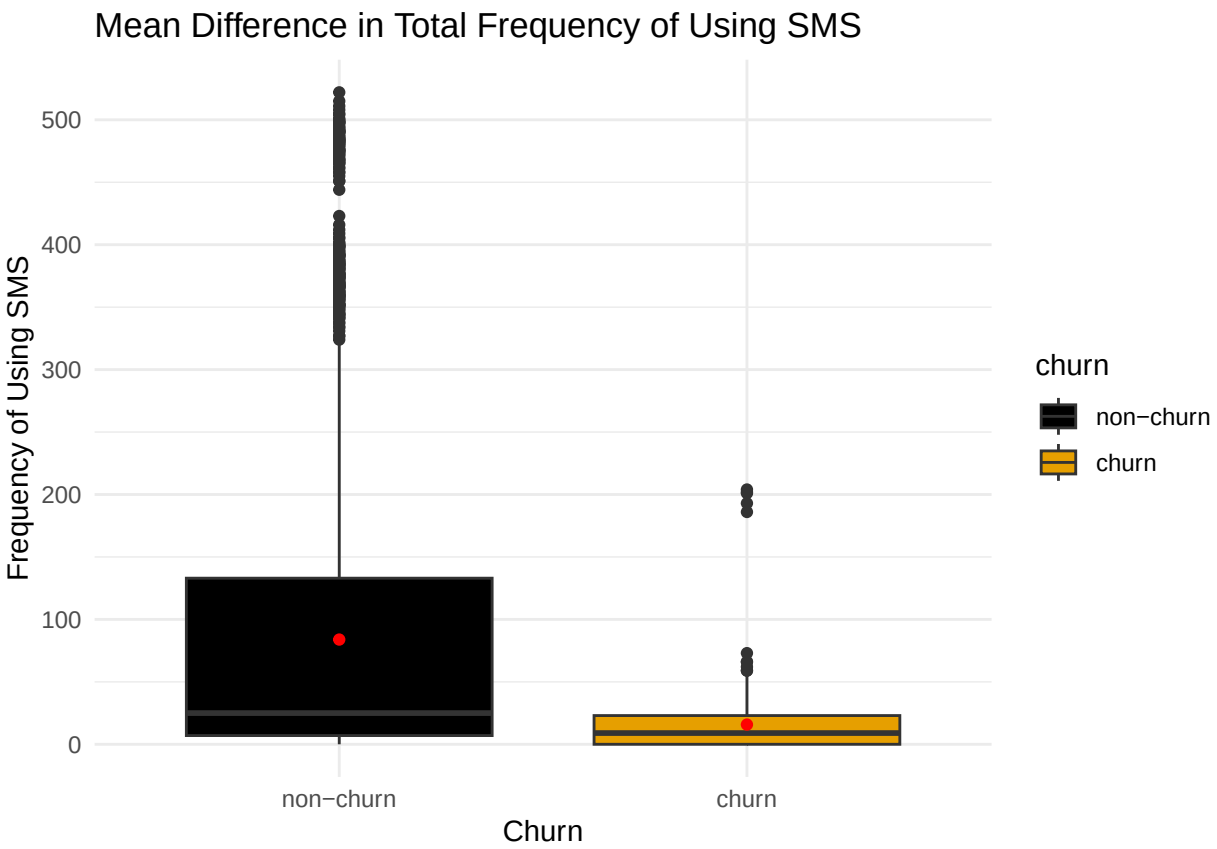


Figure 2: Average of using SMS between churn and non-churn

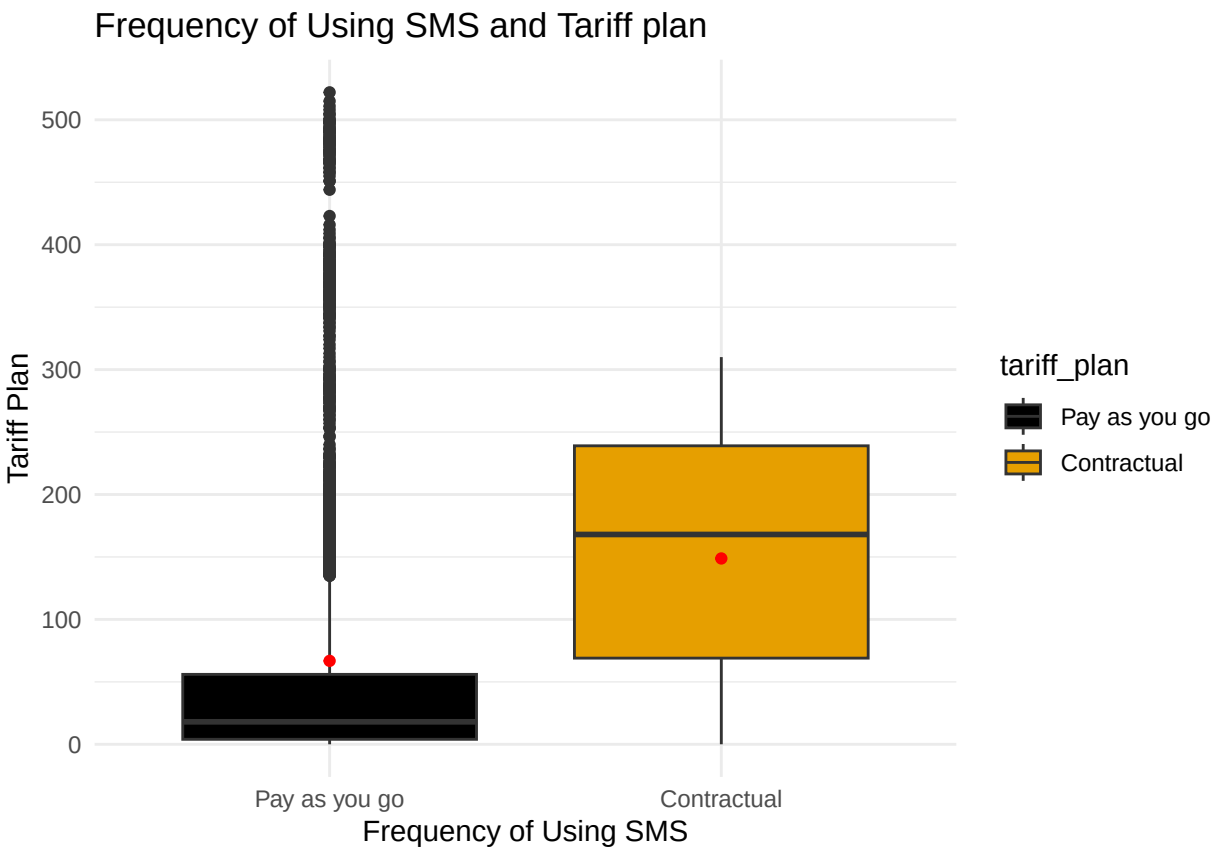


Figure 3: Average of using SMS and tariff plan

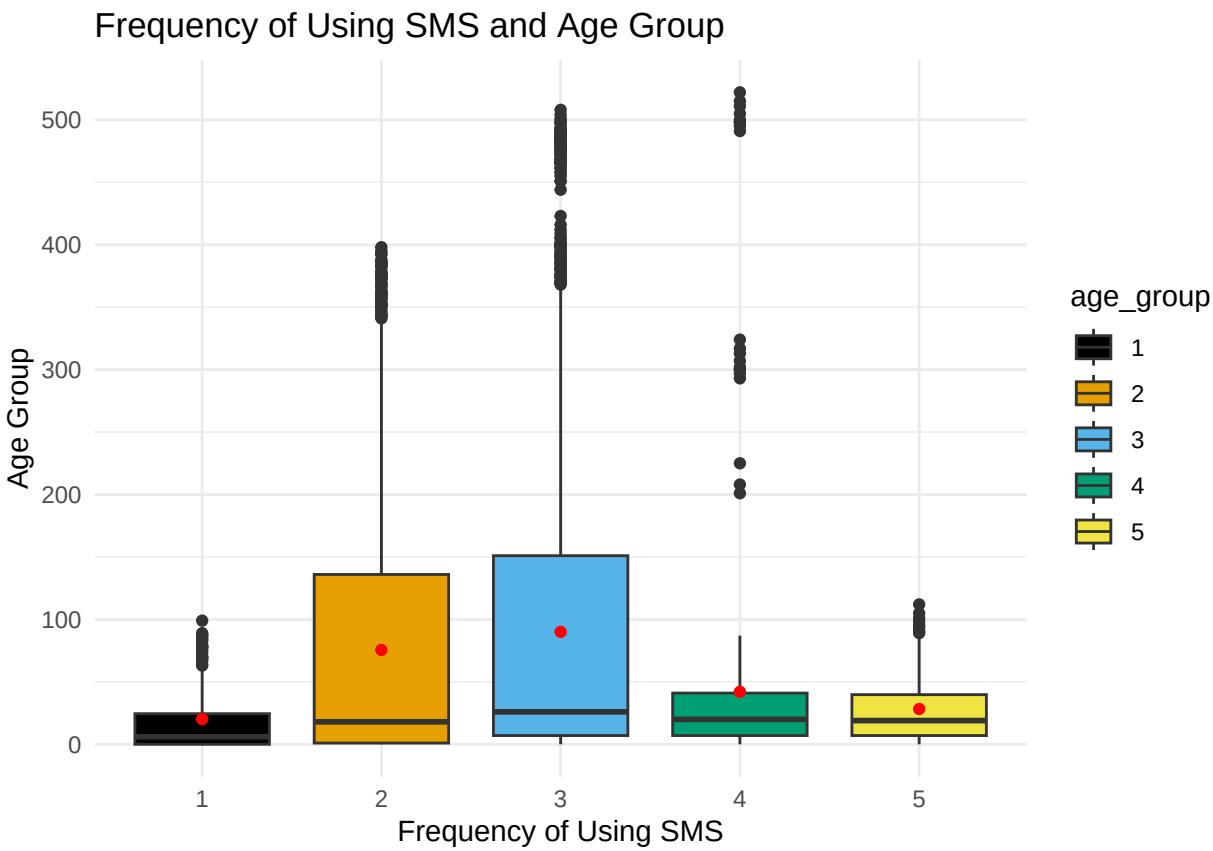


Figure 4: Frequency of using SMS and age group

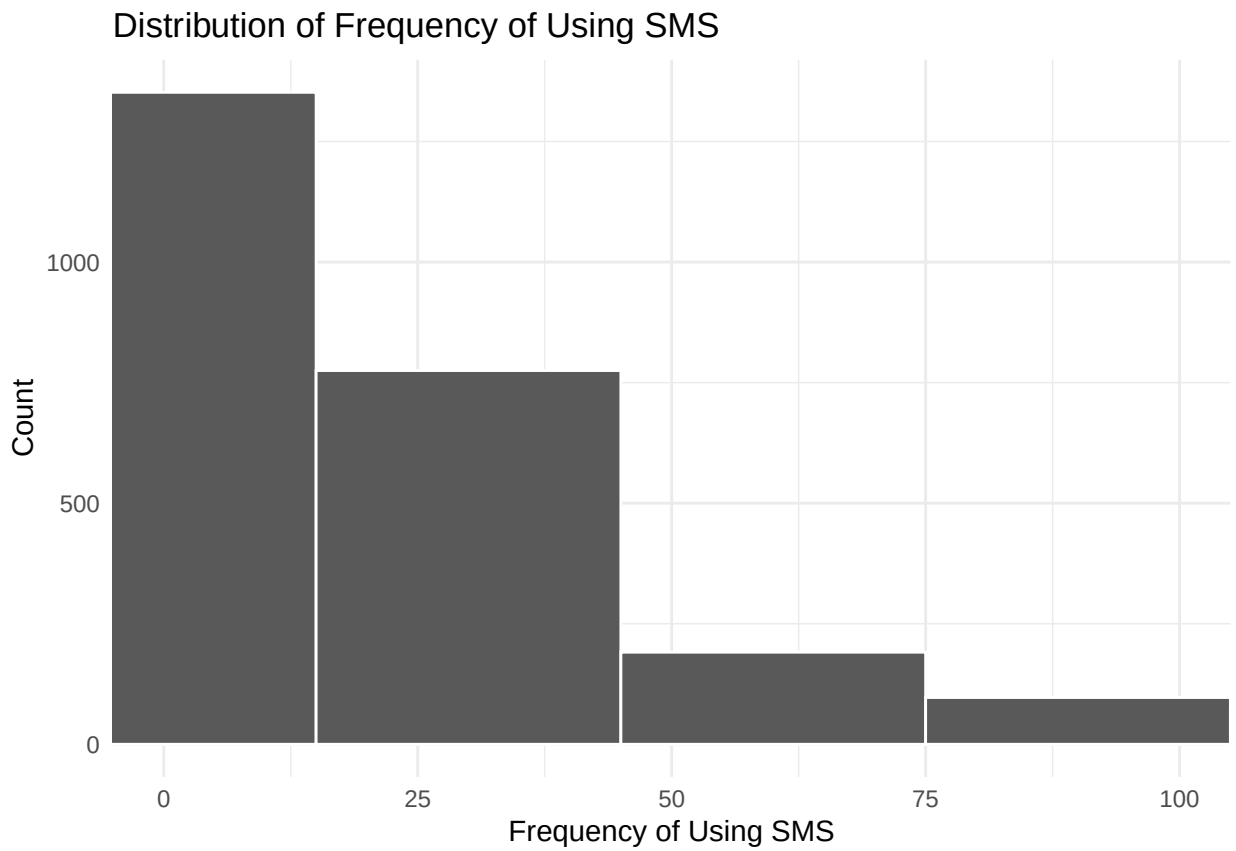


Figure 5: Distribution of Frequency SMS

- **Could there be a difference in SMS usage?** Figure 3 compares the average SMS usage between churning and non-churning customers and shows a noticeable difference in average usage.
- **Tariff plan and SMS usage** In Figure 4, despite some overlap between plans in average SMS usage, there is still a difference, suggesting a potential link between plan and churn, but deeper analysis is required to determine if plan is a significant factor.
- **Age, usage and churn** As we have seen in Figure 5, there appears to be a relationship between usage frequency, SMS usage, age groups (as in this figure) and the decision of customers to either leave or stay. Further analysis is needed here to confirm whether the company needs to improve its messaging system, tariff plans and perhaps tailor its marketing strategy more towards certain age groups.
- **Zero SMS usage?**. In Figure 6 we can see that the most common value is zero, which raises questions such as:

1. Is it plausible that SMS usage is zero for most customers?
2. The figure is skewed to the right, indicating non-normality, what could be the real value?

The zero usage of messages could happen for some customers, so dropping the values will not do much here. But since the data is randomly selected, we can use bootstrapping to see if there is a meaningful difference in the statistical analysis part.

Table 5: Summary statistics for non-churn customers

Variable Name	mean	sd	p0	p25	p50	p75	p100
call_failure	7.66	7.15	0	1.00	6.00	12.00	36.00
subscription_length	32.66	8.39	3	29.00	35.00	38.00	47.00
charge_amount	1.08	1.60	0	0.00	0.00	2.00	10.00
seconds_of_use	5014.22	4312.74	0	1819.00	3530.00	6892.50	17090.00
frequency_of_use	76.98	58.50	0	32.00	63.00	104.00	255.00
frequency_of_sms	83.87	118.81	0	7.00	25.00	133.00	522.00
distinct_called_numbers	25.58	17.39	0	12.00	23.00	36.00	97.00
age	31.07	9.15	15	25.00	30.00	30.00	55.00
customer_value	535.51	536.21	0	142.06	268.07	864.55	2165.28

Table 6: Summary statistics for churn customers

Variable Name	mean	sd	p0	p25	p50	p75	p100
call_failure	7.48	7.83	0	0.00	5.00	11.00	34.00
subscription_length	31.89	9.47	3	31.00	35.00	37.00	45.00
charge_amount	0.23	0.62	0	0.00	0.00	0.00	4.00
seconds_of_use	1566.63	1539.20	0	318.00	1182.00	2391.50	6123.00
frequency_of_use	29.13	26.32	0	6.00	25.00	45.50	100.00
frequency_of_sms	15.80	23.52	0	0.00	9.00	23.00	204.00
distinct_called_numbers	12.39	10.87	0	2.00	10.00	20.00	48.00
age	30.64	6.89	25	25.00	30.00	30.00	55.00
customer_value	124.81	129.43	0	38.38	96.84	181.32	987.26

Table 7: The correlation between frequecy of use and charge amount

churn	cor
non-churn	0.35
churn	0.13

- The summary in table 5 and 6 shows a difference in means between customers who stayed in the company for example the mean for frequency of using **SMS** is higher in customers who stayed.

Statistical analysis (What We Discovered?)

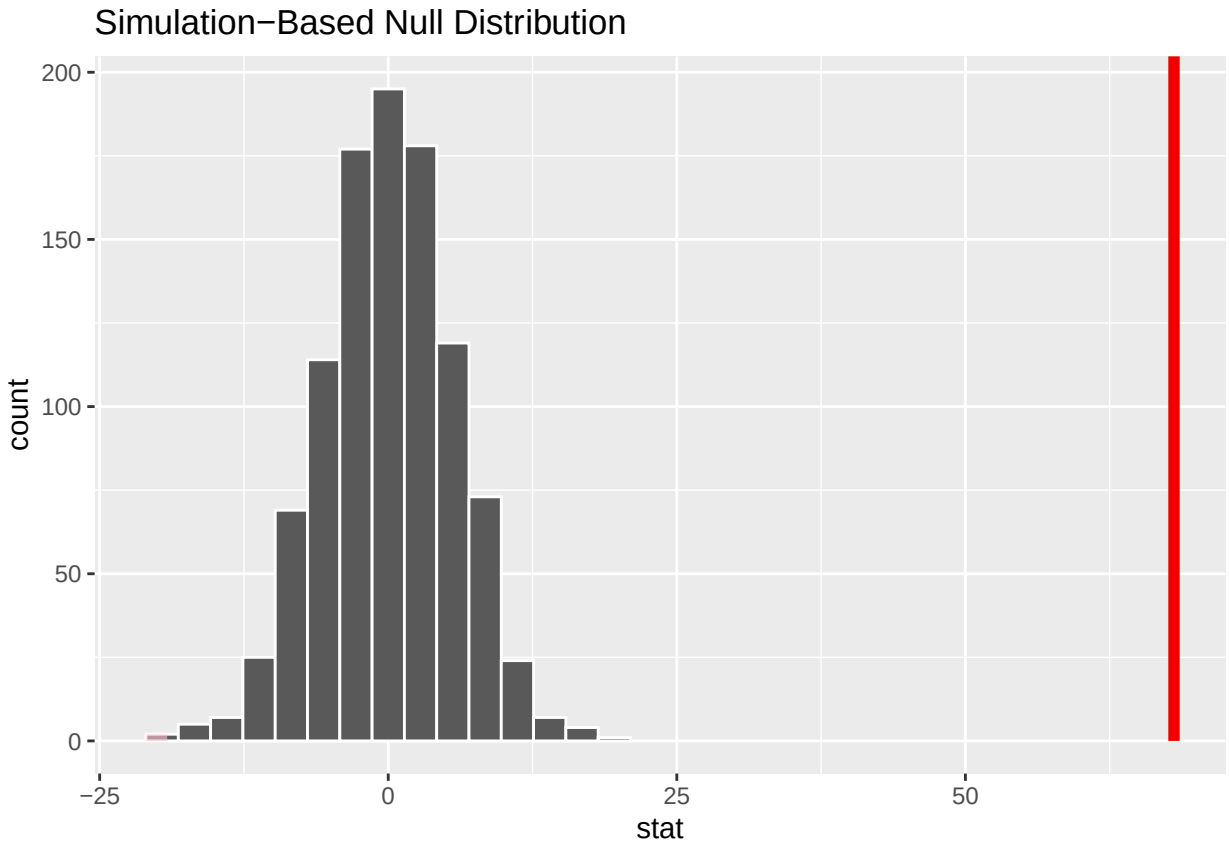
- **Assessing the significance in differences**
- Conducting hypothesis test here compare the difference, we set our level of rejection (i.e meaning that the risk of rejecting the null hypothesis when it is true) to be 5%.

Figure 3 illustrates the difference in frequency of SMS usage between churn and non churn customers.Before drawing any conclusions, we will check if the difference is statistically discernible (i.e significant).

As this will inform us about the customer behavior.

The Testing Framework :

1. **Null hypothesis**$\{H_0\}$ There is no difference in mean of frequency of SMS usage between churn and non churn customers.
2. **Alternative hypothesis**$\{H_0\}$ There is a difference in mean of SMS usage between the churn and non churn customers.



Here the the p-value 0 that i.e (the compatibility of data and the null hypothesis), we can conclude that we have a convincing evidence to reject the null hypothesis.

Indicating that there are real differences between churn and non-churn in frequency of SMS usage.

- To quantify the difference, we will construct a confidence interval for the difference in **SMS usage** :

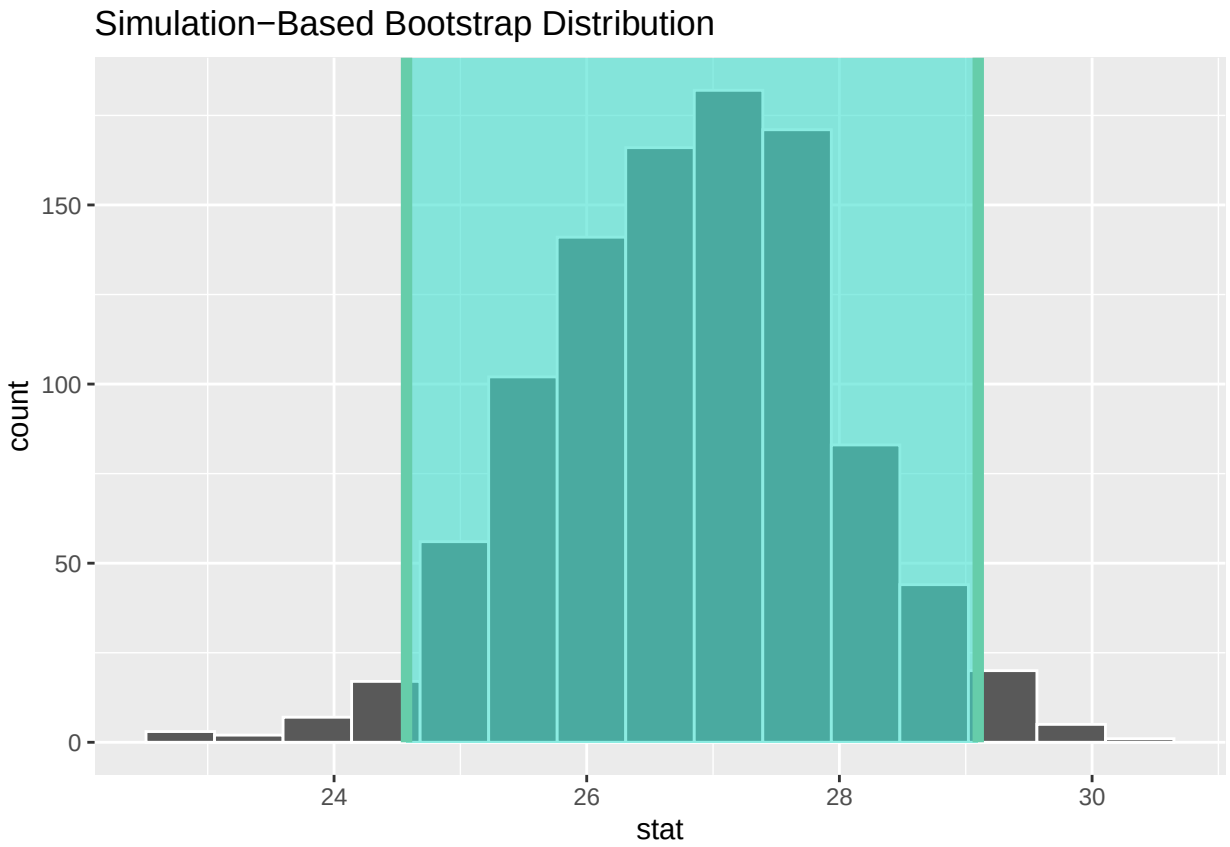


Table 8: 95% level of confidence in the range of differences in churn and non-churn customers

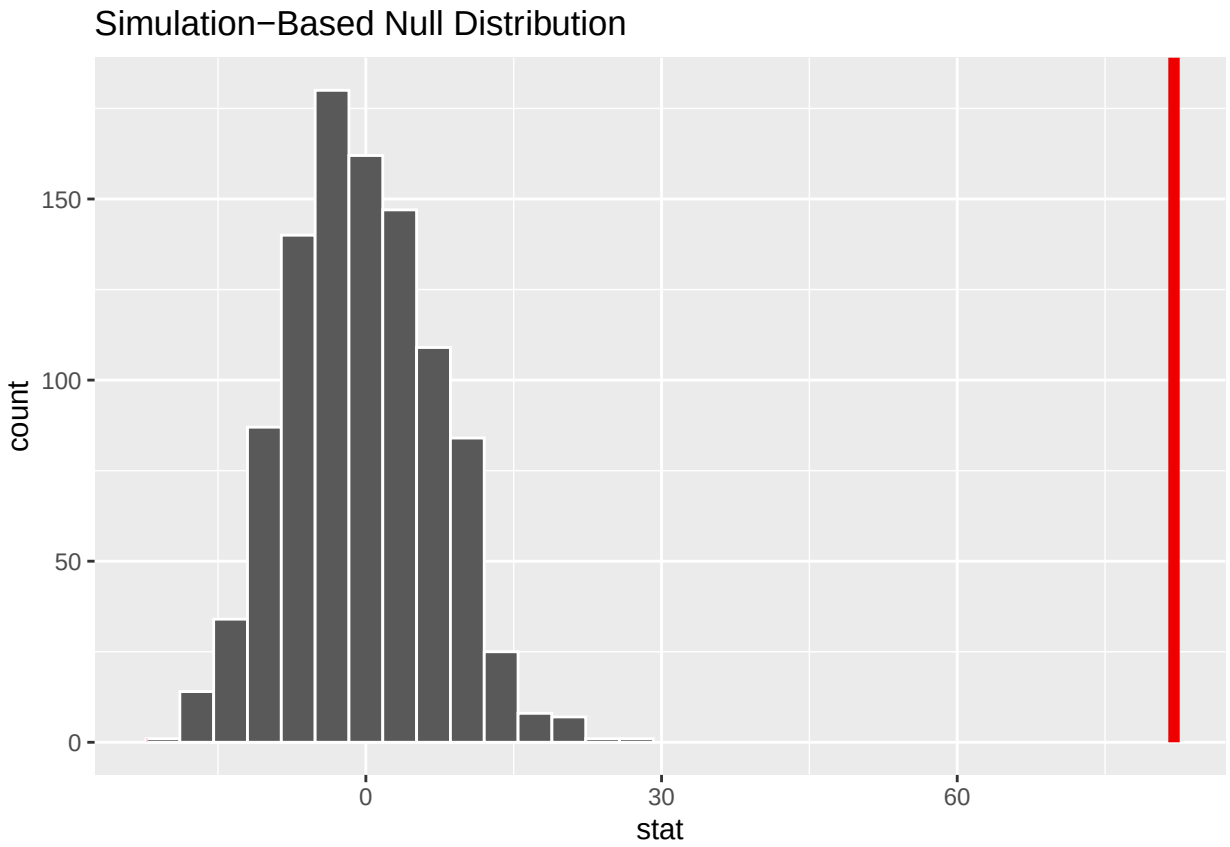
lower_ci	upper_ci
24.57	29.1

- **Assessing the Differences in Tariff Plans**

Figure 4 is showing a difference in frequency of SMS usage between tariff plans before continuing we will check if the difference is statistically discernible (i.e significant).

To do so we will conduct a hypothesis test :

1. **Null hypothesis**$\{H_0\}$ There is no difference in mean of frequency of SMS usage between the two tariff plans.
2. **Alternative hypothesis**$\{H_A\}$ There is a difference in mean of SMS usage between the two tariff plans.



- Since the p-value is 0, we can say that we have a convincing evidence to reject the null hypothesis.

Indicating that there are real differences between tariff plans in frequency of SMS usage.

An implication for the company might be asking what can make their messaging system more attractive.

Even though there are limitation to this analysis that needs to be considered, we will discuss it later.

Predictive analytics

Knowing in advance which group or subgroup of customers are likely to leave, is our aim here in which we are going to build a model step by step :

1. Fit the model: We use logistic regression for the predictive model and apply backward selection to choose variables with significant p-values.

Table 9: A table for logistic regression based-model

term	estimate	std_error	statistic	p_value	conf_low	conf_high
(Intercept)	-0.43	0.03	-14.64	0	-0.49	-0.38
call_failure	0.17	0.00	62.25	0	0.17	0.18
charge_amount	-0.75	0.02	-41.12	0	-0.79	-0.72
frequency_of_sms	-0.01	0.00	-68.63	0	-0.01	-0.01

frequency_of_use	-0.03	0.00	-55.96	0	-0.04	-0.03

- Table 9 summarizes the logistic regression model after backward selection, showing only the variables with significant p-values.

2.For the logistic regression model to provide valid results, certain assumptions must be satisfied :

Independence: The data has been collected randomly, so we assume sufficiency in this condition.

Linearity: Which for this condition to work linear relationship between logit and predictor variables needs to exist:

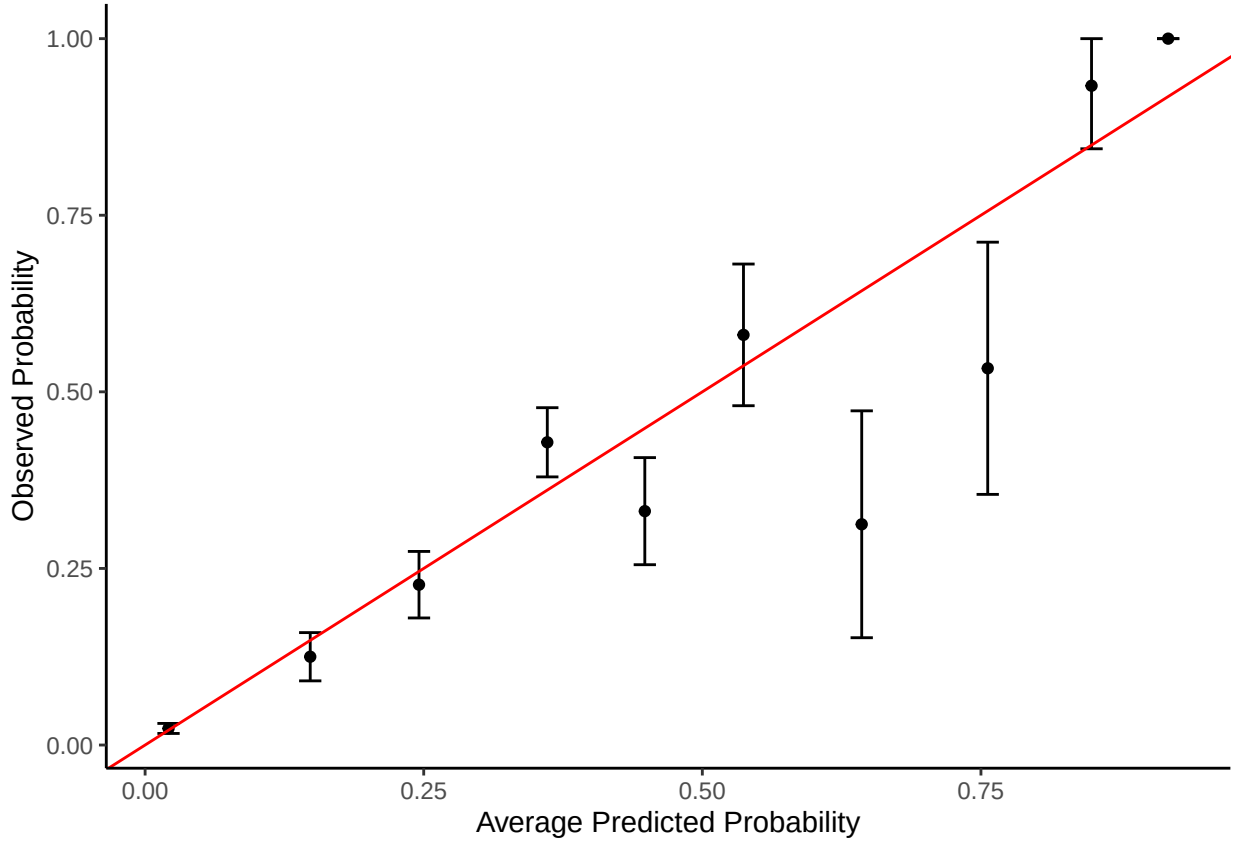


Figure 6: Model assumptions

The plot compares predicted probabilities (x-axis) with observed probabilities (y-axis). The points should cluster around the red diagonal line (representing perfect predictions). This indicates that the model's linearity assumption holds since the observed probabilities match predicted values closely.

3. After checking the conditions, we check the accuracy of the model (i.e how much the model can explain the churn variable).

To do so we will use 5-fold cross-validation

The accuracy of the model is **0.86**.

- We can evaluate the model in another way using confusion matrix :

Table 10: Confusion matrix

	non-churn	churn
churn	77	123
non-churn	2578	372

The table there is telling true negatives, positives, and false negatives and positives(i.e how many predictions that our model got it right and it was about 0.86).

But this accuracy might be **misleading** because the model was trained and tested on the same data set.

We will take the same steps but training and testing on different data sets.

Table 11: More accurate evaluation on training and testing set

	non-churn	churn
churn	4	26
non-churn	1675	285

The accuracy rate here is 0.85 and Miss-classification rate is 0.15.

4. Model predictions. We plug in some values to the model and get the probability of churning :

For example when (call failure is 8, 10), (charge_amount 1, 3), (frequency of sms 10, 14) and (frequency_of_use 82, 44) :

Table 12: Probability of churning for certain customers

x
0.06
0.07

As the task at hand is only to do prediction and not to interpret coefficients, a model that suffers from high multicollinearity will likely lead to unbiased predictions of the response variable. So multicollinearity is likely to not cause any substantial problems.

But we should be careful about **extrapolations**. In other words, just because our model supports a linear relationship doesn't mean that relationship holds for values outside our range. Predictions for such values is far away from the actual range often aren't accurate.

Conclusion

Our model, with an accuracy rate of 0.85, offers valuable predictive power, meaning the company can identify customers at risk of churn and intervene with tailored retention strategies. While some miss-classifications occurred, the model still provides strong actionable insights.

Limitations

- **Observational Data**: Since this is an observational study, causality cannot be established. While the model can predict the likelihood of churn, it does not identify the exact causes behind customer decisions to leave. This means any results should be interpreted as a correlation, not causation.
- **Potential Biases**: The data is limited to one telecom company, we may not be able to generalize to other sectors or regions.

Future improvment

1. Experimentation for Causal Inference.
2. Considering more external Factors.