Waterman Workspaces Customer Churn Analysis: Impact of CRM Cases

Bilal Raja

2024-10-02

Contents

1 Abstract **2**

2 Background and Motivation **2**

3 Objectives and Significance **2**

4 Methodology **3**

4.1 Data Cleaning and Joining 3

4.2 Creating the Final Data Subset 4

4.3 Exploratory Data Analysis 5

4.4 Logistic Regression Modeling 9

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
##
## Attaching package: 'kableExtra'
##
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
##
##
##
## Attaching package: 'gridExtra'
##
```

```
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
##
## Loading required package: lattice
##
##
## Attaching package: 'caret'
##
##
## The following object is masked from 'package:purrr':
##
##     lift
```

1 Abstract

This report discusses the customer churn analysis for Waterman Workspaces by using the customers' membership data, together with the data from the Customer Relationship Management (CRM) system to derive insights for investigating the factors that may potentially influence the customers' churn behaviour. Factors such as case escalations, frequency of cases, and time taken to resolve a case were investigated using logistic regression modelling. The purpose of our analysis is to use the findings to guide retention strategies and improve services of the organization.

2 Background and Motivation

Waterman Workspaces uses a membership subscription model to run their business which are defined as their 'products'. Some examples of the products include casual hire, team membership, part time membership, and suite among others. Due to an intense market competition in the domain, the organization found it pertinent to investigate customer retention on a precautionary basis in order to uncover an aspect of their business' health. Two of the most crucial areas were analyzing the customers' attendances and foot traffic and case analysis using the data from the CRM. The area to investigate was the CRM for me where the analysis focused on whether the case patterns - that include the number of cases and their resolution time - and their escalation statuses affect the likelihood of churn.

Through the CRM system, Waterman Workspaces keeps a record of the numerous interactions that take place with their customers. However, given the data, it is unclear how the information from the data source correlate with churn. Thus, it is pertinent to investigate the dynamics between the service quality provided by the organization and churn in order to formulate actionable insights and enhance customer satisfaction.

3 Objectives and Significance

The objectives and significance of the report include:

- Identifying the key drivers for churn: Assess the variables that influence that lead to customers churning. The variables investigated in the report include the number of cases, their resolution time, case escalations, and whether different sites of Waterman Workspaces have greater churns. Accurate analysis gives birth to formulating targeted strategies.

Table 1: List of Variables in the 'wmcases' Dataset

Variables	Variables
(Do Not Modify) Case	(Do Not Modify) Row Checksum
(Do Not Modify) Modified On	Case Title
Account Number (Customer) (Account)	Customer
Case Age (Days)	Follow Up By
Case Note	Site
Priority	Status Reason
Is Escalated	Modified On
Case Age	Case Number
Case Type	Satisfaction
Sentiment Value	Service Level
SLA	Severity
Status	Description
Subject	(Do Not Modify) Case

Table 2: List of Variables in the 'memberships' Dataset

Variables	Variables
Product Name	Created On
Billing End Date	Billing Start Date
Created By	Lessee
Location	Status
Status Reason	Total Monthly Lease
Accounting Code	Lease Products
Product Category	Account Number (Lessee) (Account)
Account Category (Lessee) (Account)	Industry (Lessee) (Account)

- Modelling and quantifying the impact: Determining how the potential drivers affect churn likelihood via logistic regression modeling.
- Develop actionable insights for long-term success: Provide data-driven recommendations to the organization that are concise and understood with ease for the business for retention strategies. Fostering informed decision-making leads to enhanced overall business performance.

4 Methodology

4.1 Data Cleaning and Joining

The data was provided the host organization in the form of 2 primary datasets; CRM Cases data (`wmcases`) and membership data (`memberships`).

The CRM Cases data output is seen below where we can see the variables that come along with it followed by the variables in the memberships data (see Tables 1 and 2).

Upon consultation with the relevant personnel in the organization, the variables and observations that were not needed were removed from our analysis. Moreover, since the key variable was the *Account Number* of the customers, due diligence was made to ensure blanks and non-customer observations were removed as well. When the irrelevant observations and variables were removed, the 2 datasets were joined using a left join by using the `Account Number` as the key variable. The left join was essential as it retained all records

Table 3: Variables in the cases joined Dataset

Variables	Variables
Account Number	Customer
Case Number	Case Title
Case Age	Case Age (Days)
Is Escalated	Follow Up By
Case Note	Site
Case Type	Status.x
Description	Subject
Created On	Billing End Date
Billing Start Date	Created By
Status Reason.y	Total Monthly Lease
Lease Products	Product Category
Account Category (Lessee) (Account)	Account Number

of the CRM data and the matching records of the memberships data so that every case lodged in the CRM system was included in the analysis even if some memberships data for customers was missing. In theory, this approach factors in customers that may have recently joined in or left already.

4.2 Creating the Final Data Subset

The merged data produced a large dataset with an increasing number of variables that needed to be cleaned further. As before, potential `NA` values were catered for and further nonessential variables and observations were removed. The resulting dataset had the following variables listed in Table 3:

Next, a vector was created to define churn statuses that includes

- Inactive - Customer Cancelled
- Off-boarding
- Inactive

The mentioned attributes were selected upon consultations and discussions with the supervisor.

Using the defined churn statuses, a new subset of the data was formed, namely `customer_features`. The cases were grouped by their account numbers and summarized to count the number of cases per account and a similar process was followed to summarize by the average case age as well to provide insights into a the typical duration of cases for each customer. In addition, the median was taken for the case ages over a mean to handle the skewness of the data. In such scenarios, mean is sensitive to extreme data points which could provide an untrue picture of the time it takes to resolve a case. On the other hand, the median takes the ***50th percentile*** (the central point). Moreover, `churned` was created as a binary variable. The logic behind making it binary was that either a customer churns or does not churn and this case, `1` meant churned and `0` meant not churned.

Additionally, `product_category` was created which retained the first product associated with the account to reflect the main product used by the customer. The initial reasoning behind including the product categories was to analyze whether different categories had varying churn rates. A similar approach was taken to incorporate the `site` variable to subset to investigate a sites of the organizations have varying churn rates that could point towards service quality issues. The joined dataset was further filtered to include another binary variable `Is Escalated` that was created to provide numeric representation to of case escalation. Here, when a case is marked "Yes", the final value is converted to `1` and `0` otherwise. Finally, steps were taken to ensure that the `product_category` and `site` variables were converted to factor.

The final variables used for our analysis and their explanation are as follows:

Table 4: Variables in the customer features Dataset

Variables
Account Number
num_cases
case_age
churned
product_category
site
is_escalated

- `Account Number`: A unique identifier of a customer.
- `num_cases`: The total number of cases per customer.
- `case_age`: The median age in days taken to resolve a case.
- `churned`: A binary indicator that shows whether a customer churned or did not churn.
- `product_category`: The type of membership subscribed by the customer.
- `site`: The site or the location of the customer of Waterman.
- `is_escalated`: A binary indicator that shows whether a case raised by the customer was escalated.

4.3 Exploratory Data Analysis

4.3.1 Ensuring the Data is Clean and Structured

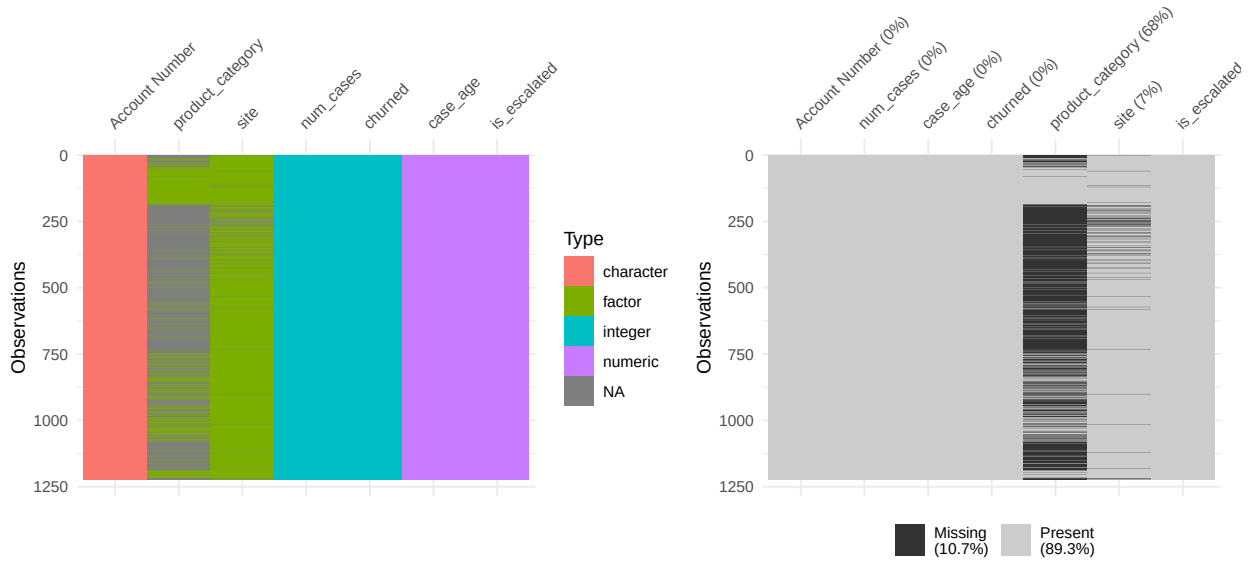


Figure 1: Comparison of dataset structure and missing values.

While we are able to what type the variables are, Figure 1 shows that there are a significant number of missing observations in `product_category`. Due to a significantly high proportion of missing values of the

product categories, the variable was not included in our analysis as the missing values would distort our analysis and provide little to no meaningful analysis.

4.3.2 Summary Description

Table 5: Custom Summary for Selected Variables

num_cases	case_age	churned	is_escalated	product_category	site
Min. : 1.000	Min. : 0.00	Min. :0.0000	Min. :0.0000	Suite :112	Chadstone :364
1st Qu.: 1.000	1st Qu.: 3.00	1st Qu.:0.0000	1st Qu.:0.0000	Access Membership : 95	Caribbean Park :316
Median : 2.000	Median : 12.00	Median :0.0000	Median :0.0000	Part Time Membership: 65	Narre Warren :157
Mean : 4.246	Mean : 39.07	Mean :0.3186	Mean :0.1046	Dedicated Desk : 49	Eastland :115
3rd Qu.: 4.000	3rd Qu.: 40.00	3rd Qu.:1.0000	3rd Qu.:0.0000	Anchor Suite : 18	Sixty Four on Victor: 85
Max. :171.000	Max. :1605.00	Max. :1.0000	Max. :1.0000	(Other) : 51	(Other) :106
NA	NA	NA	NA	NA's :834	NA's : 81

Table 5 shows the overall summary of the subset. It illustrates that a customer raises 4.246 issues on average and it takes an average of 39 days to resolve the issue. Both `num_cases` and `case_age` have medians lower than the mean that signal positive skewness which will be visualized next. Moreover, the table provides further information on the total number of each type of product a customer has subscribed to and the umber of customers at each site.

4.3.3 Distribution of Case Frequency and Case Age (with outliers)

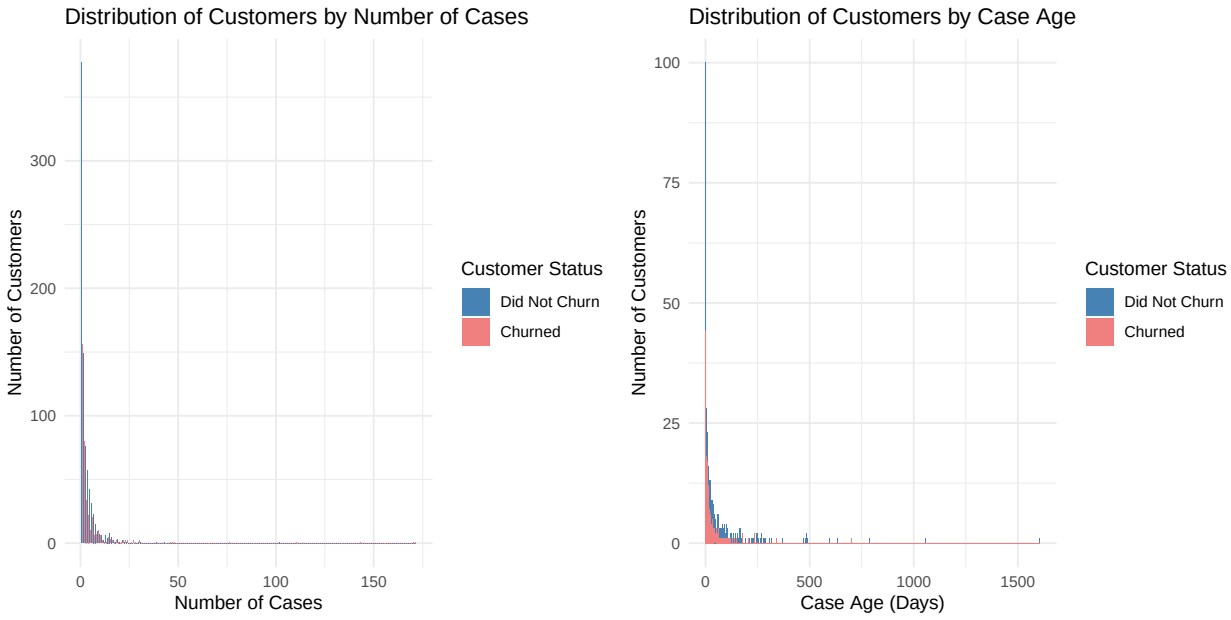


Figure 2: Outlier Distributions

Figure 2 reveals high positive skewness for the number of cases and average case age. These are such extreme data points that intuition dictated to investigate further. However, it was recommended to not include extreme outliers for business purposes. Simultaneously, Figure 2 offered an insight into how the data may be managed currently by the organization, hinting at problems that need to be addressed.

Since it was recommended to remove the outliers, the interquartile (IQR) method was employed for their removal in order to paint a clearer picture for our analysis. Using the interquartile method was preferred to any other method because of its robustness and by using the middle 50% of the data, our analysis is not influenced by the extreme values observed. Hence, the IQR method allows for preserving the underlying distribution of the data without any trade-off of the data integrity and models.

For self reflection purposes, it was realized that using the median for the average case ages previously did not remove the outliers.

4.3.4 Churn Status Comparison

Before visualizing the number of cases and the case resolution times, we first review a holistic comparison of customers churning against the customers who did not churn from the final cleaned subset of the data. Figure 3 below illustrates that There are 314 customers who churned and 676 customers who did not churn.

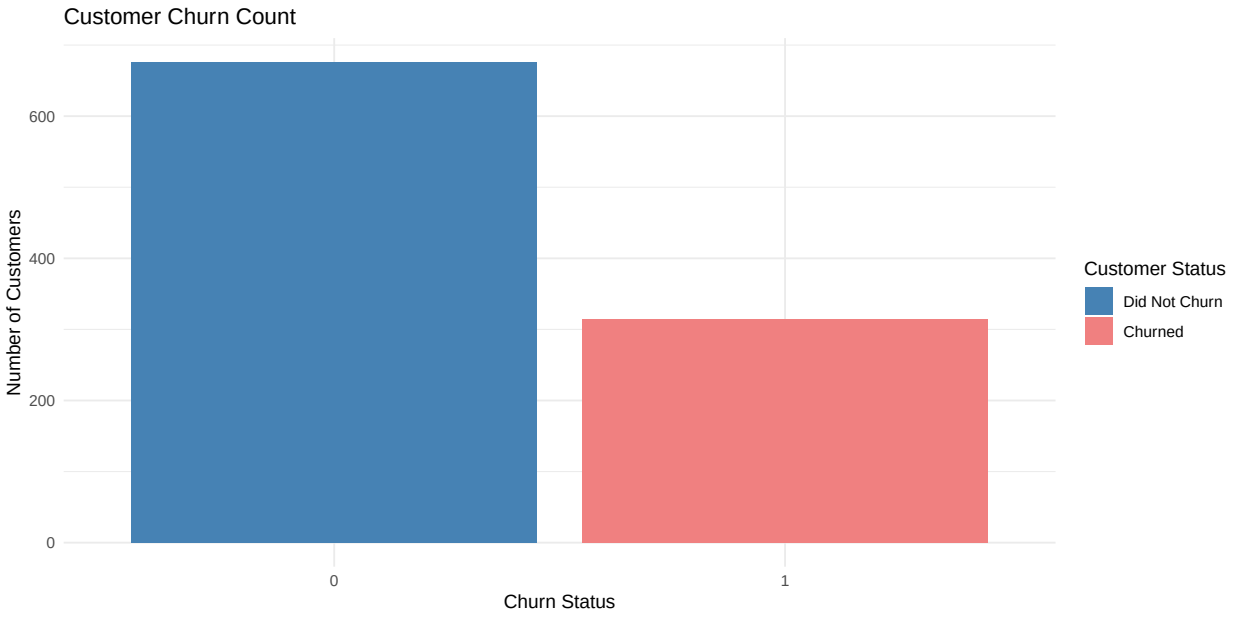


Figure 3: Analyzing Customers Churning

4.3.5 Distribution of Case Frequency and Case Age (cleaned)

Next, we plot the histograms to analyze the distributions from the cleaned data.

Figure 4 shows number of customers against the number of cases. The customers that churned and those who did not churn are segregated by separate colours for ease in visual comparison.

The figure suggests that the distribution of customers by their number of cases is positively skewed which means that there are more customers that faced issues less frequently. Moreover, it shows that as the number of cases increase, the number of customers churning decreases significantly. While this may seem counter-intuitive, it also shows that there decreasing number of customers that face more problems. In this case, it is more important to analyze the proportion of customers that churned versus the customers that did not churn given the number of cases. Even so, the number of customers that leave are still lower than the customers that do not leave at every bin of number of cases.

Figure 5 shows number of customers against the average case age. The customers that churned and those who did not churn are segregated by separate colours for ease in visual comparison.

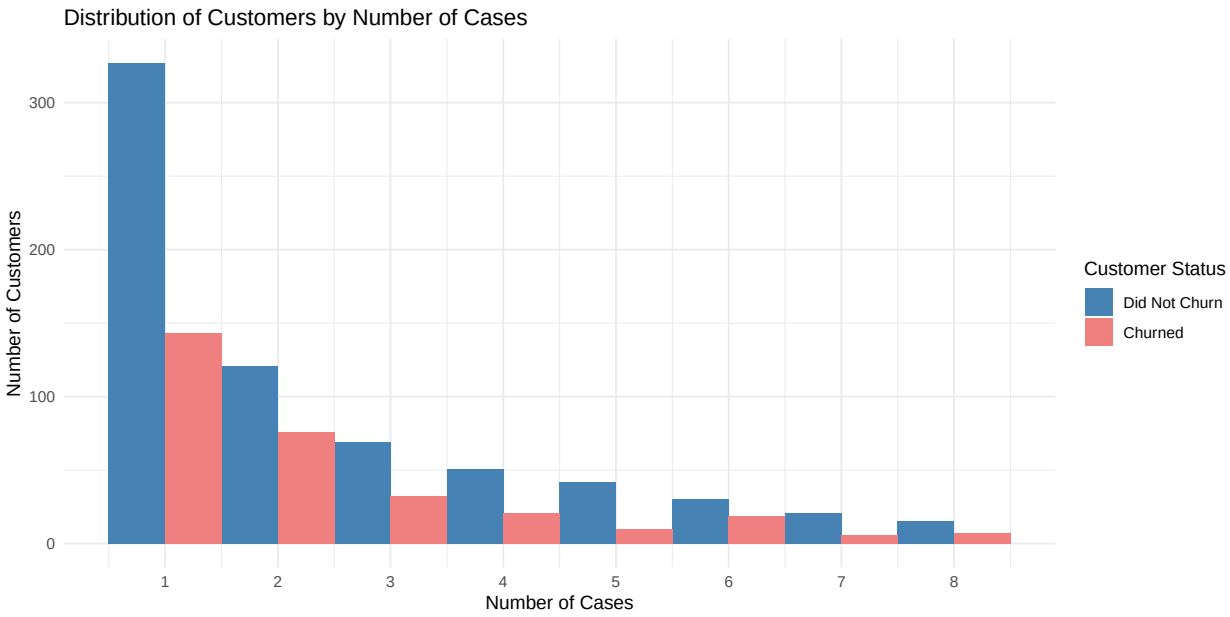


Figure 4: Number of Cases Distribution

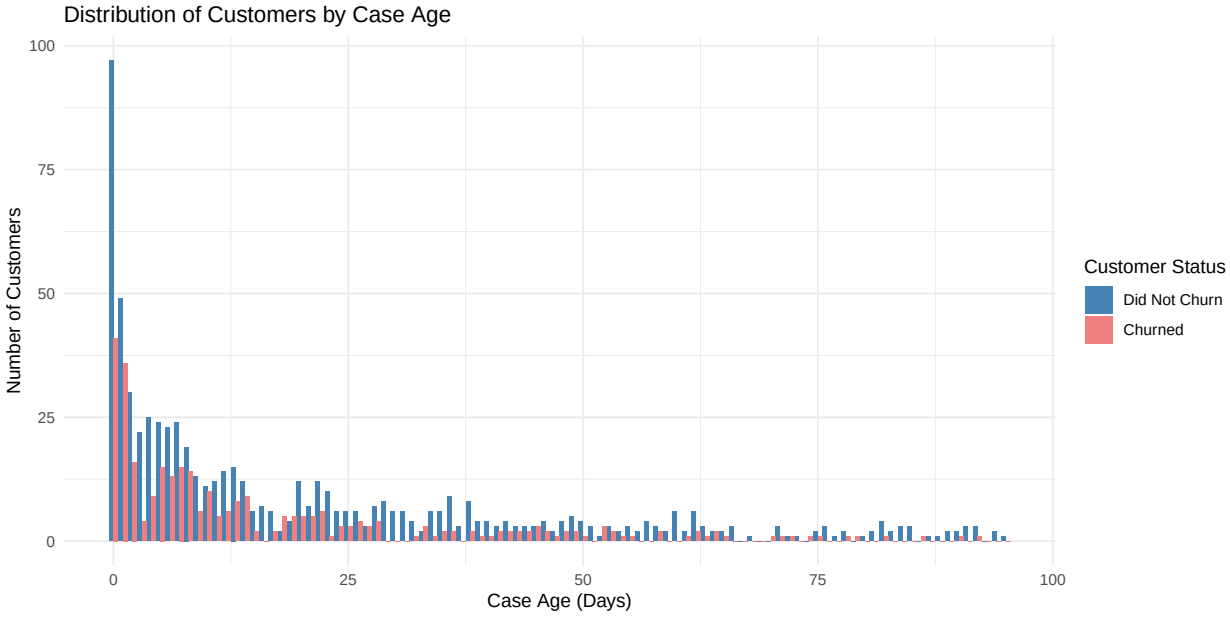


Figure 5: Case Age Distribution

Table 6: Correlation Matrix

	num_cases	case_age	is_escalated	churned
num_cases	1.0000000	-0.0120297	0.0550789	-0.0186413
case_age	-0.0120297	1.0000000	-0.1504204	-0.0875657
is_escalated	0.0550789	-0.1504204	1.0000000	0.4771995
churned	-0.0186413	-0.0875657	0.4771995	1.0000000

The figure suggests that the distribution of customers against the case age is also positively skewed which means that there are more customers that had cases resolved mostly sooner rather than later. Moreover, it shows that as the case resolution time increases, the number of customers churning decreases overall. This too seems counter-intuitive and also shows that there are less customers that face longer case resolution times.

4.3.6 Correlation

The correlation matrix in Table 6 produces some interesting points. The correlations between all variables are very weak and inconclusive apart from `is_escalated`. While it was assumed that as the number of cases and case resolution times increase, the impact on churn would increase as well. However, this positive relationship is not observed. On the other hand, cases being escalated have a positive correlation with customers churning albeit having a weak correlation of 0.4771995

4.4 Logistic Regression Modeling

For the purpose of our analysis, we applied logistic regression to model the relationships between the explanatory variables and the outcome of customers churning. According to Peng et al. (2002) generally, logistic regression is well suited for describing and testing hypotheses about relationships between a categorical outcome variable and one or more categorical or continuous predictor variables. In the simplest case of linear regression for one continuous predictor X and one dichotomous outcome variable Y, the plot of such data results in two parallel lines, each corresponding to a value of the dichotomous outcome. Thus, the binary nature of our outcomes encourage us to use logistic regression.

Moreover, the interpretability of logistic regression is fairly easier as it indicates how the predictors impact the outcome (in terms of log-odds). Since we take into account the probabilistic outcome, it helps us to identify at-risk customers and develop retention strategies. While more complex models could have been possible to apply, such as decision trees, logistic model was preferred to avoid unnecessary complications and provide easy to interpret results for the business audience.

Finally, since the goal of our task was to explore relationship between the predictors and customers churning, the data was not split into training and testing since the goal is exploratory in nature and not prediction.

The logistic regression model can be expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{num\_cases} + \beta_2 \text{case\_age} + \beta_3 \text{is\_escalated} + \beta_4 \text{product\_category} + \beta_5 \text{site}$$

Where: - p is the probability of customer churn. - β_0 is the intercept. - $\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients for the predictors: number of cases, case age, escalation status, and site, respectively.

Once the log-odds are calculated, they are converted to a probability using the following equation:

$$p = \frac{e^{\beta_0 + \beta_1 \text{num\_cases} + \beta_2 \text{case\_age} + \beta_3 \text{is\_escalated} + \beta_4 \text{site}}}{1 + e^{\beta_0 + \beta_1 \text{num\_cases} + \beta_2 \text{case\_age} + \beta_3 \text{is\_escalated} + \beta_4 \text{site}}}$$

This equation produces the log-odds output between 0 and 1, which represents the probability of the customer churning.

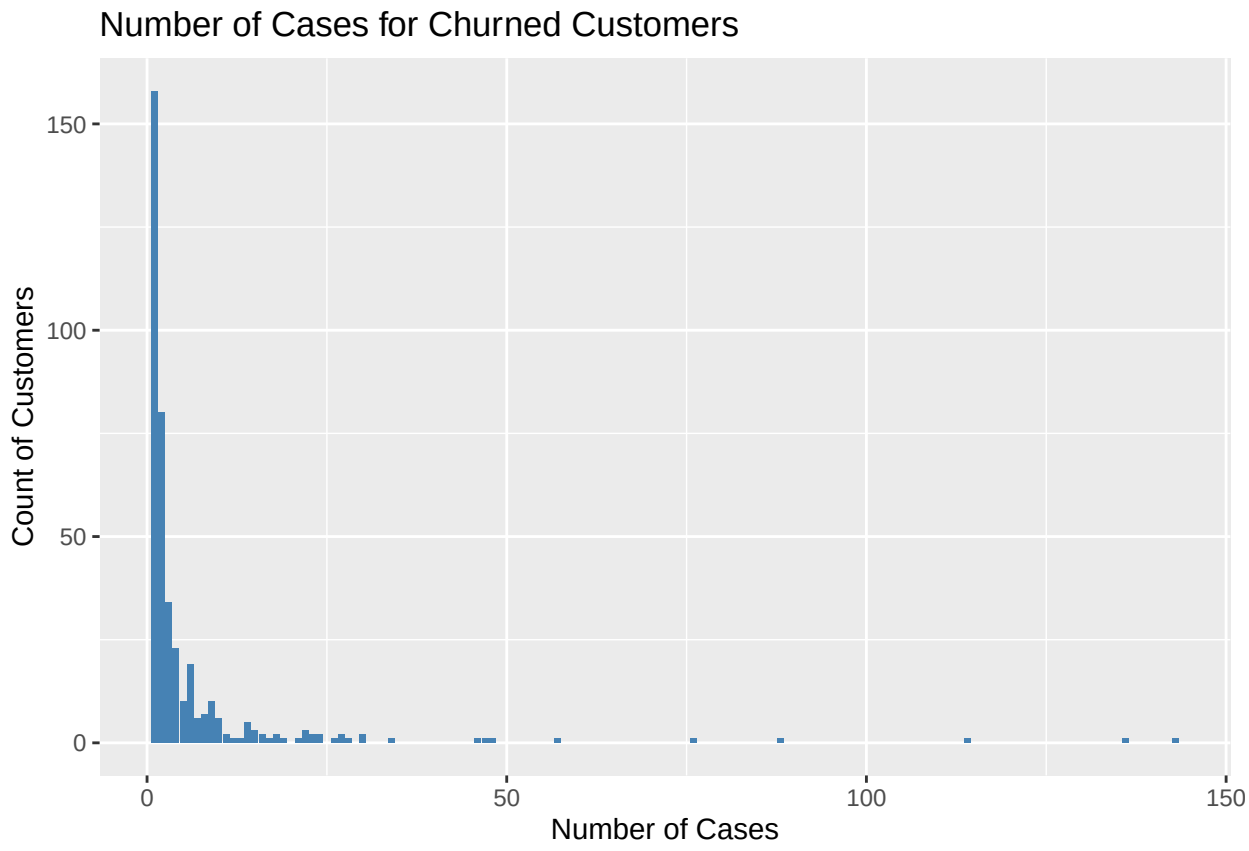
```
##
## Call:
## glm(formula = churned ~ num_cases + case_age + is_escalated +
##     site, family = binomial, data = customer_features_clean)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.837e+01  4.945e+02  -0.037    0.970
## num_cases                   3.730e-02  5.021e-02   0.743    0.458
## case_age                    1.708e-03  3.773e-03   0.453    0.651
## is_escalated                4.609e+00  6.647e-01   6.934 4.09e-12 ***
## siteBundoora                1.967e+01  4.945e+02   0.040    0.968
## siteCamberwell              1.948e+01  4.945e+02   0.039    0.969
## siteCaribbean Park          1.728e+01  4.945e+02   0.035    0.972
## siteCasey Corporate Centre  1.593e+00  1.767e+03   0.001    0.999
## siteChadstone               1.671e+01  4.945e+02   0.034    0.973
## siteEastland                1.891e+01  4.945e+02   0.038    0.969
## siteGENERAL ENQUIRY         3.485e+01  2.450e+03   0.014    0.989
## siteNarre Warren            1.676e+01  4.945e+02   0.034    0.973
## siteRichmond                1.754e+01  4.945e+02   0.035    0.972
## siteSixty Four on Victor    1.384e+01  4.945e+02   0.028    0.978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1169.46  on 913  degrees of freedom
## Residual deviance:  793.44  on 900  degrees of freedom
##   (76 observations deleted due to missingness)
## AIC: 821.44
##
## Number of Fisher Scoring iterations: 15
```

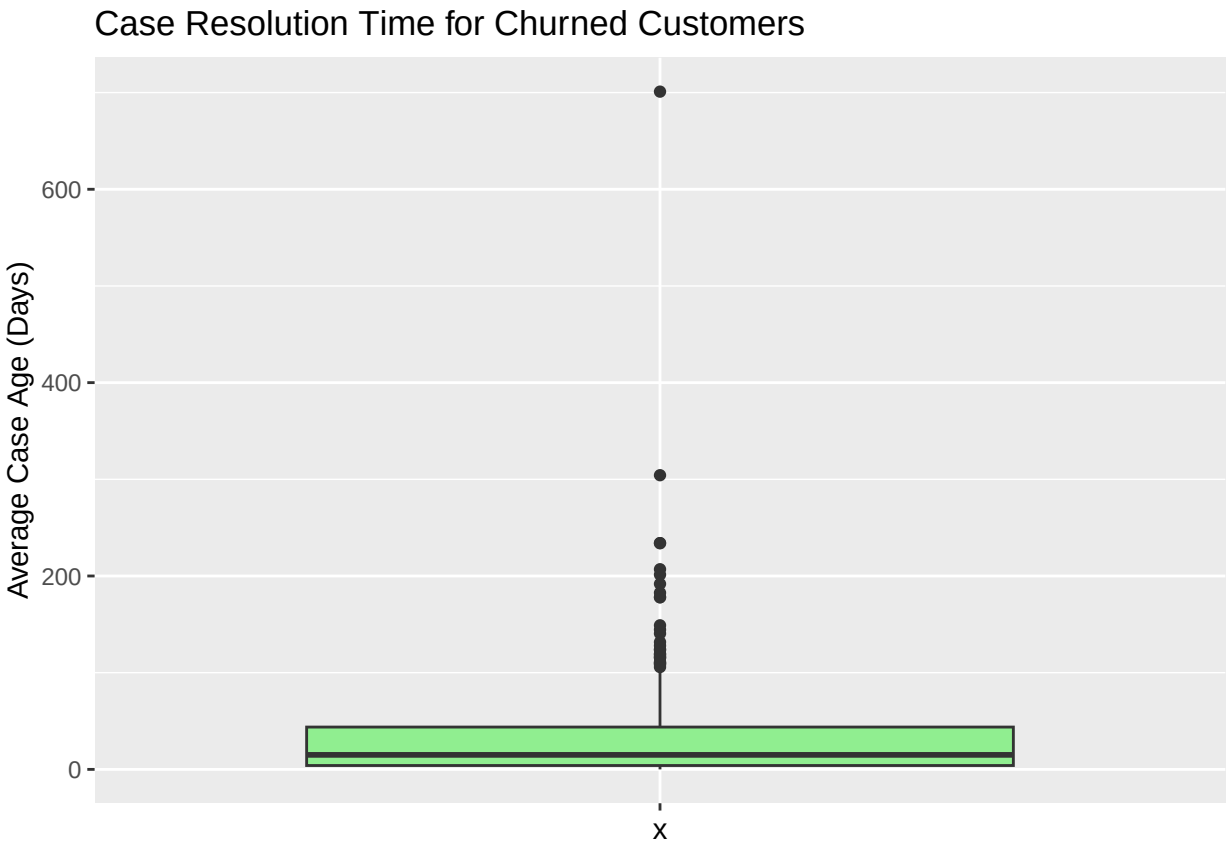
```
##
## Call:
## glm(formula = churned ~ num_cases + case_age + is_escalated +
##     site + num_cases_escalated + case_age_escalated, family = binomial,
##     data = customer_features_clean)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.802e+01  5.057e+02  -0.036    0.972
## num_cases                   3.978e-02  5.062e-02   0.786    0.432
## case_age                    1.623e-03  3.790e-03   0.428    0.668
## is_escalated                4.775e+00  1.192e+00   4.005 6.21e-05 ***
## siteBundoora                1.931e+01  5.057e+02   0.038    0.970
## siteCamberwell              1.912e+01  5.057e+02   0.038    0.970
## siteCaribbean Park          1.693e+01  5.057e+02   0.033    0.973
## siteCasey Corporate Centre  1.231e+00  1.770e+03   0.001    0.999
## siteChadstone               1.635e+01  5.057e+02   0.032    0.974
## siteEastland                1.856e+01  5.057e+02   0.037    0.971
```

```
## siteGENERAL ENQUIRY         3.450e+01  2.452e+03   0.014    0.989
## siteNarre Warren            1.641e+01  5.057e+02   0.032    0.974
## siteRichmond                1.720e+01  5.057e+02   0.034    0.973
## siteSixty Four on Victor    1.344e+01  5.057e+02   0.027    0.979
## num_cases_escalated        -1.465e-01  3.348e-01  -0.437    0.662
## case_age_escalated          2.258e-02  5.925e-02   0.381    0.703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1169.46  on 913  degrees of freedom
## Residual deviance:  793.14  on 898  degrees of freedom
##   (76 observations deleted due to missingness)
## AIC: 825.14
##
## Number of Fisher Scoring iterations: 15

##           df      AIC
## log_model 14 821.4432
## int_model 16 825.1422

##           df      BIC
## log_model 14 888.8928
## int_model 16 902.2275
```




Case Resolution Time for Churned Customers
Average Case Age (Days)
600
400
200
0
x