Insights on Customer Churn

Chatterbox Telco Pvt Ltd

CS 3120 Introduction to Data Science

Name : Kajanan Selvanesan

Index Number : 190287R

Date : 01.05.2022

Problem overview

Customer churn occurs when a company loses consumers for various reasons, including bad service and lower rates elsewhere. It is one of the most significant and challenging concerns for telecommunication companies, credit card companies, cable service providers, and other businesses.

Mr. Williams, CEO of Chatterbox Telecom Pvt Ltd in the Banana Republic, wants to analyse this customer churn in his company. Since acquiring new customers costs more than retaining existing ones, analysing customer churn and deriving valuable insights that would be useful for Mr. Williams to make strategic decisions to improve customer retention.

Dataset description

To analyse customer churn and derive valuable insights, we were provided with a dataset that contains the package type and usage details of a customer and whether they left Chatterbox or not. In particular, the dataset consists of 19 predictor variables and one target variable.

Variable Name	Data Type	Attribute Type	Description
customer_id	Integer	Categorical Nominal	Customer identification number.
account_length	Integer	Metric Discrete	Number of months the customer has been with the current telco provider.
location_code	Integer	Categorical Nominal	Customer location code.
international_plan	String	Categorical Nominal	If the customer has an international plan or not.
voice_mail_plan	String	Categorical Nominal	If the customer has a voice-mail plan or not.
number_vm_messages	Integer	Metric Discrete	Number of voice-mail messages.
total_day_min	Float	Metric Continuous	Total minutes of day calls.
total_day_calls	Integer	Metric Discrete	Total number of day calls.
total_day_charge	Float	Metric Continuous	Total charge of day calls.
total_eve_min	Float	Metric Continuous	Total minutes of evening calls.
total_eve_calls	Integer	Metric Discrete	Total number of evening calls.

total_eve_charge	Float	Metric Continuous	Total charge of evening calls.
total_night_minutes	Float	Metric Continuous	Total minutes of night calls.
total_night_calls	Integer	Metric Discrete	Total number of night calls.
total_night_charge	Float	Metric Continuous	Total charge of night calls.
total_intl_minutes	Float	Metric Continuous	Total minutes of international calls.
total_intl_calls	Integer	Metric Discrete	Total number of international calls.
total_intl_charge	Float	Metric Continuous	Total charge of international calls.
customer_service_calls	Integer	Metric Discrete	Number of calls to customer service.
Churn	String	Categorical Nominal	If the customer left or not (target variable).

We were provided with two datasets: the training and testing datasets. The training dataset contains the variables mentioned above with one extra column. The testing dataset contains the variables discussed above, excluding the Churn, with two additional columns. However, all these columns have some quality issues.

Let's see the dataset descriptions of both,

	count	mean	std	min	25%	50%	75%	max
customer_id	2321.0	2161.000000	670.159309	1001.00	1581.000	2161.00	2741.0000	3321.00
account_length	2319.0	101.400172	40.044985	1.00	74.000	101.00	127.0000	232.00
location_code	2321.0	473.470918	42.011853	445.00	445.000	452.00	452.0000	547.00
number_vm_messages	2318.0	7.557377	14.250001	-202.00	0.000	0.00	14.0000	51.00
total_day_min	2320.0	182.718103	73.332822	-179.90	144.000	180.35	221.0000	2283.90
total_day_calls	2318.0	105.324418	221.100535	-1.00	87.000	102.00	115.0000	10700.00
total_day_charge	2316.0	30.961524	9.830271	-25.60	24.480	30.60	37.5900	60.96
total_eve_min	2318.0	203.511734	115.552100	-103.30	165.925	202.40	236.4000	5186.40
total_eve_calls	2317.0	100.125162	20.536224	-80.00	87.000	101.00	114.0000	170.00
total_eve_charge	2313.0	17.123130	4.327327	0.00	14.180	17.21	20.0900	30.83
total_night_minutes	2319.0	209.543467	408.066120	23.20	167.350	201.10	235.0500	19700.00
total_night_calls	2316.0	87.641192	12.737232	33.00	79.000	90.00	98.0000	105.00
total_night_charge	2316.0	9.436710	18.656075	1.04	7.530	9.05	10.5825	900.15
total_intl_minutes	2319.0	10.247736	2.795472	-9.30	8.600	10.30	12.0000	18.30
total_intl_calls	2318.0	4.439172	2.461172	0.00	3.000	4.00	6.0000	20.00
total_intl_charge	2316.0	2.773364	0.733526	0.00	2.320	2.78	3.2400	4.94
customer_service_calls	2320.0	1.651724	1.429166	0.00	1.000	1.00	2.0000	9.00
Unnamed: 20	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Train dataset

	count	mean	std	min	25%	50%	75%	max
customer_id	1500.0	4071.500000	433.157015	3322.00	3696.7500	4071.500	4446.2500	4821.00
account_length	1500.0	101.042000	39.454167	1.00	73.0000	100.000	127.2500	243.00
location_code	1498.0	475.508678	43.035587	445.00	452.0000	452.000	547.0000	547.00
number_vm_messages	1499.0	7.805871	13.376356	0.00	0.0000	0.000	18.0000	50.00
total_day_min	1497.0	184.498798	56.977595	2.60	144.6000	184.100	222.1000	345.30
total_day_calls	1497.0	100.085504	20.531492	-85.00	87.0000	100.000	113.0000	157.00
total_day_charge	1496.0	31.314693	10.014655	-46.48	24.6425	31.300	37.8575	59.36
total_eve_min	1498.0	203.982443	51.534663	42.50	168.3250	203.850	238.1750	363.70
total_eve_calls	1500.0	100.034000	19.994950	44.00	87.0000	100.000	113.0000	168.00
total_eve_charge	1491.0	17.321415	4.366784	3.61	14.2700	17.330	20.2100	30.91
total_night_minutes	1497.0	201.957448	50.607989	-207.40	168.4000	203.500	236.2000	381.90
total_night_calls	1498.0	119.414553	10.920603	105.00	111.0000	117.000	125.0000	175.00
total_night_charge	1498.0	9.093071	2.223011	1.97	7.5800	9.140	10.6475	17.19
total_intl_minutes	1498.0	10.326101	2.917300	0.00	8.4000	10.300	12.3000	20.00
total_intl_calls	1497.0	4.409486	2.538735	-5.00	3.0000	4.000	6.0000	18.00
total_intl_charge	1500.0	2.789387	0.787599	0.00	2.2700	2.795	3.3200	5.40
customer_service_calls	1499.0	1.638426	1.385127	0.00	1.0000	1.000	2.0000	9.00
Unnamed: 19	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Unnamed: 20	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Test dataset

Within the variables discussed above, 4 are categorical columns: location_code (445 or 452 or 547), international_plan (yes or no), voice_mail_plan (yes or no), and Churn (Yes or No).

Data pre-processing

Data Cleaning

There are some quality issues in both datasets. Therefore, it is essential to do data cleaning on both.

1. Detect & Remove Duplicates

   Without considering the customer_id, if you see the training dataset, there are four duplicate rows in the indexes 772, 807, 1704, and 2295. I removed those rows. In the same way, you cannot detect any duplicate rows in the test dataset.

2. Handling Missing Values, Out of Range Values & Outliers

   In both datasets, there are negative values. But no attribute gets negative values here. I replaced such values in both datasets by using their absolute values. I took this decision with the help of observations made on the datasets.

```
customer_id                0
account_length             2
location_code              0
intertiol_plan             3
voice_mail_plan            6
number_vm_messages         3
total_day_min              1
total_day_calls            3
total_day_charge           5
total_eve_min              3
total_eve_calls            4
total_eve_charge           8
total_night_minutes        2
total_night_calls          5
total_night_charge         5
total_intl_minutes         2
total_intl_calls           3
total_intl_charge          5
customer_service_calls     1
Churn                      5
Unnamed: 20             2317
dtype: int64
```

Train dataset

```
customer_id                0
account_length             0
location_code              2
intertiol_plan             4
voice_mail_plan            4
number_vm_messages         1
total_day_min              3
total_day_calls            3
total_day_charge           4
total_eve_min              2
total_eve_calls            0
total_eve_charge           9
total_night_minutes        3
total_night_calls          2
total_night_charge         2
total_intl_minutes         2
total_intl_calls           3
total_intl_charge          0
customer_service_calls     1
Unnamed: 19             1500
Unnamed: 20             1500
dtype: int64
```

Test dataset

Most of the columns contain missing values in both datasets. You can see them in the pictures below.

I removed unwanted columns (Unnamed: 19 or Unnamed: 20 that contain full of null values) from the datasets.

I used the boxplot and scatterplot to identify some outliers and out-of-range values. Initially, I replaced those values with null. After that, I dealt with the missing values. When I dealt with missing values, I also dealt with outliers and out-of-range values.

If any column in the total_day_min, total_day_call, or total_day_charge has a value of zero for a row, those values are substituted with regard to the values in the other two columns. If any of the two values for the columns listed above is zero in a row, then the third column in that row is also zero. This technique is applied to evening, night, and international calls also.

Variable Name	How Did I Handle Missing Values?
customer_id	There are no missing values.
account_length	Filled using the median.
location_code	Filled using the last valid observation.
international_plan	Filled using the most frequent value (no).
voice_mail_plan	Filled using the most frequent value (no).
number_vm_messages	If voice_mail_plan == no: number_vm_messages = 0 Else if voice_mail_plan == yes: Filled using the median of where voice_mail_plan == yes.
total_day_min (/total_eve_min/ total_night_minutes/ total_intl_minutes)	Sorted the datasets by total_day_charge (/ total_eve_charge/ total_night_charge/ total_intl_charge) then filled missing values using the linear interpolation of the variable before and after a timestamp for a missing value.
total_day_calls (/total_eve_calls/ total_night_calls/ total_intl_calls)	Sorted the datasets by total_day_min (/ total_eve_min/ total_night_minutes/ total_intl_minutes) then filled missing values using the last valid observation.
total_day_charge (/total_eve_charge/ total_night_charge/ total_intl_charge)	Sorted the datasets by total_day_min (/ total_eve_min/ total_night_minutes/ total_intl_minutes) then filled missing values using the linear interpolation of the variable before and after a timestamp for a missing value.
customer_service_calls	Filled using the most frequent value (1).
Churn	Removed the specific rows.

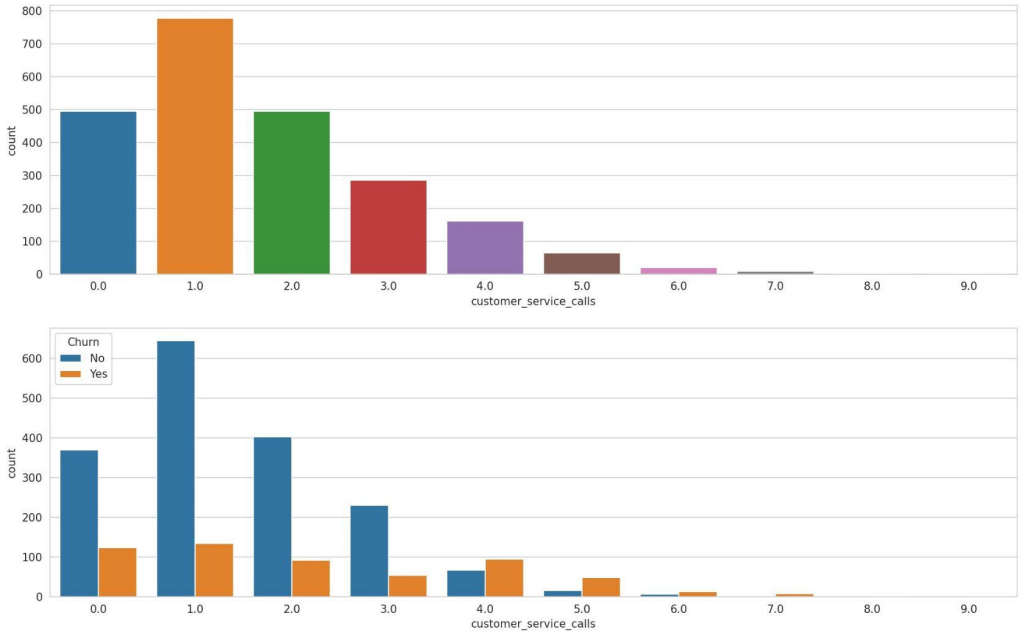
Data Transformations

1. Data Encoding

   There are four categorical columns altogether that need to be encoded. I used a dummy encoding technique for the location code, and label encoding technique for the international_plan, voice_mail_plan and Churn. When using label encoding, I used 1 for yes/Yes and 0 for no/No.

Insights from data analysis

Insight 1: If a customer has called more than three times to the customer service centre of the Chatterbox, then most probably that customer will churn.



You can notice from the picture below that the number of churned customers in each count of customer_service_calls is greater than that of unchurned customers when considering the customers who had called more than three times to the customer service centre of the Chatterbox.

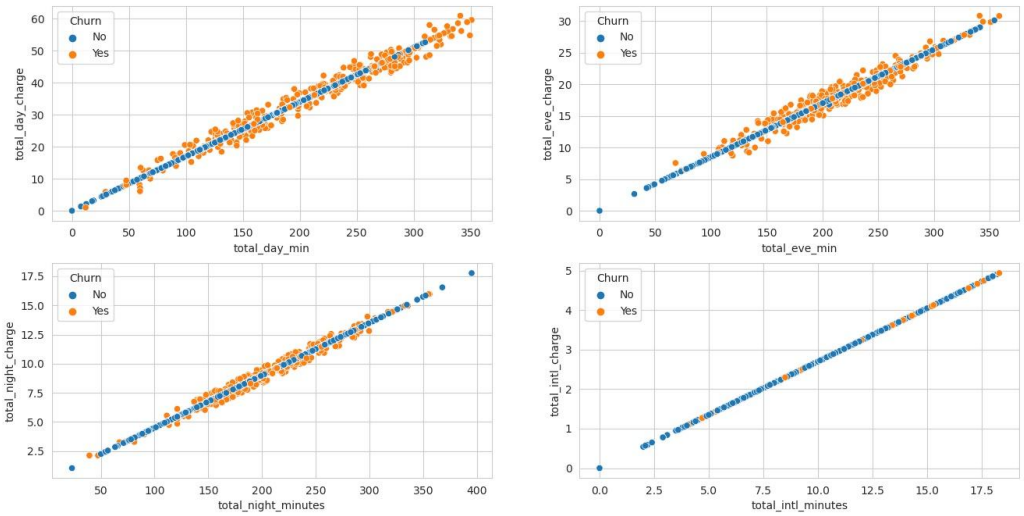
And also from the past data which is shown below you can notice that all the customers who have called more than six times to the customer service centre of the Chatterbox, have churned.

customer_id	customer_service_calls	Churn
1370	7.0	Yes
1403	7.0	Yes
1498	9.0	Yes
1605	9.0	Yes
2146	7.0	Yes
2462	8.0	Yes
2503	7.0	Yes
2627	7.0	Yes
2635	7.0	Yes
2765	7.0	Yes
2881	7.0	Yes
2920	7.0	Yes
3191	8.0	Yes

To reduce customer becoming churn,

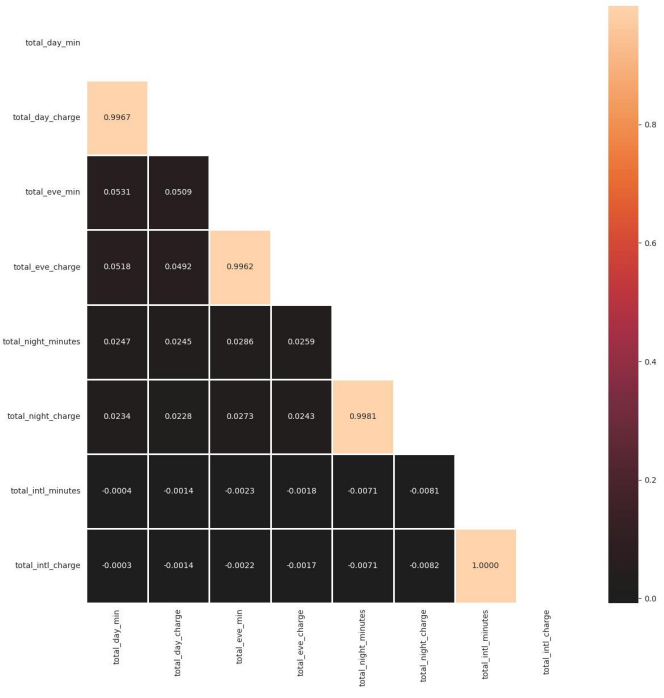
1. The company needs to fulfil the customer's needs who call most time. And take extra care of those customers.
2. Another reason for becoming churn may be the company's customer service gives poor customer service. Therefore improving the customer service quality may reduce the churns.

Insight 2: There are possibilities for a customer to churn when Chatterbox charges an unusual charge per minute from the customer during the day or evening, or night.

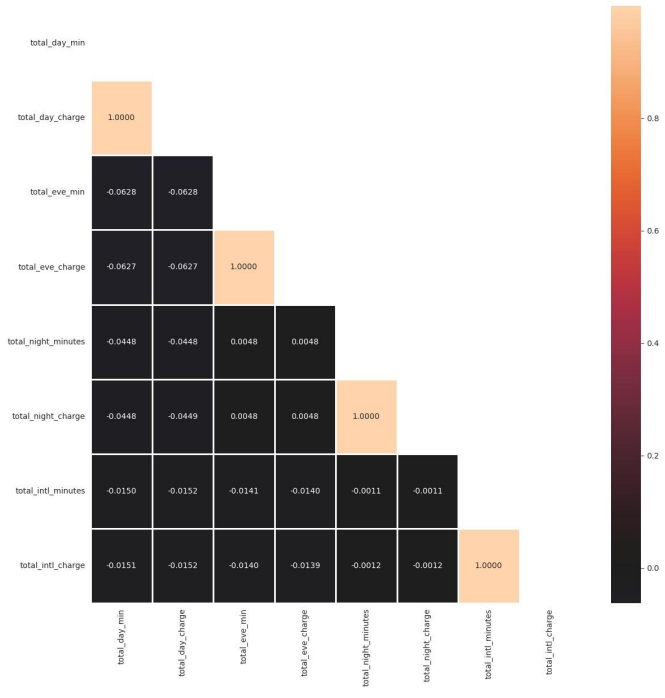


The below picture shows the relationship between the total minutes and the total charge for any specific period. Here, you can notice that the most probably churned customers are not in the plotted straight line. But that's not the case for unchurned customers.

I calculated the correlations between the above minutes-charge pairs in two cases: with and without churned customers. The results are shown in the pictures below. When using the dataset with churned customers, the correlations between the above minutes-charge pairs are closest to the one. But when using the dataset without churned customers, the correlations between the above minutes-charge pairs are exactly one.

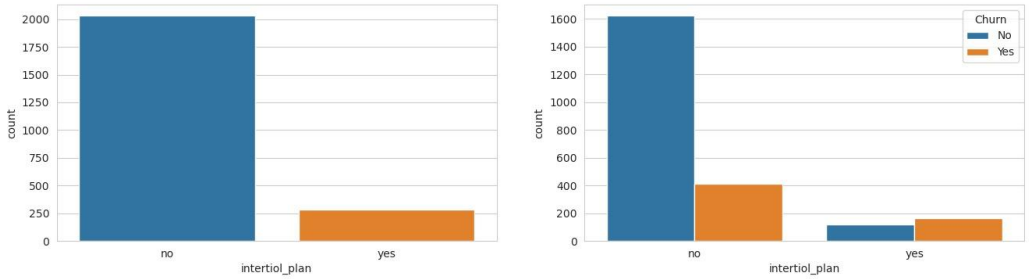


With Churn

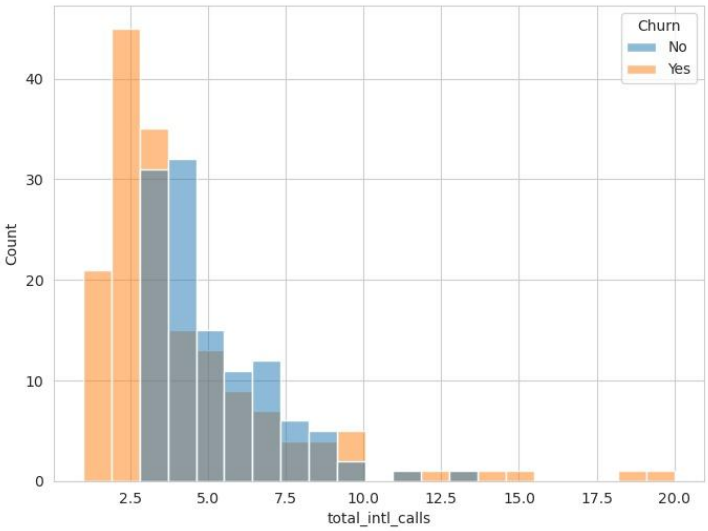


Without Churn

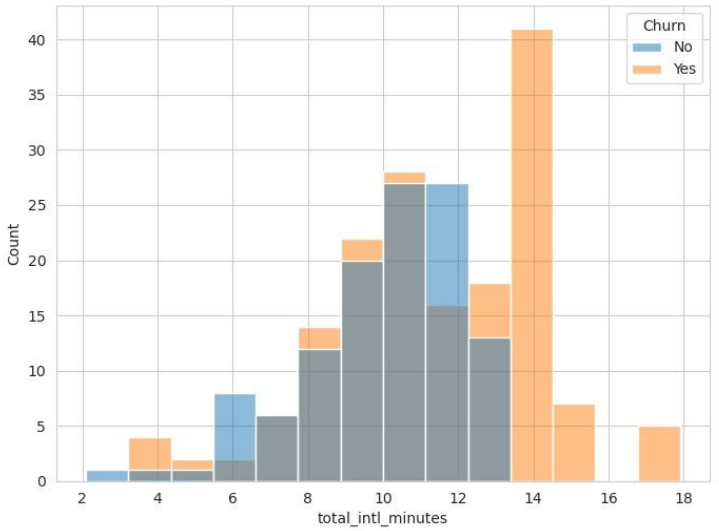
Insight 3: If a customer has international_plan, then most probably that customer will churn.



You can notice from the picture below that the number of churned customers is greater than the number of unchurned customers when considering the customers who have international_plan because you may have been providing a poor international_plan.

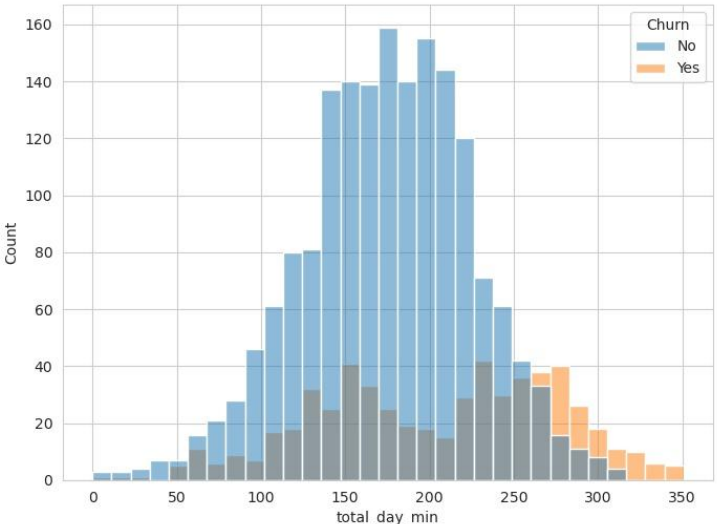


When only considering the customers who have international_plan, many of the customers have churned with a few international calls and some of the customers have churned with a lot of international calls. It's shown in the above picture.

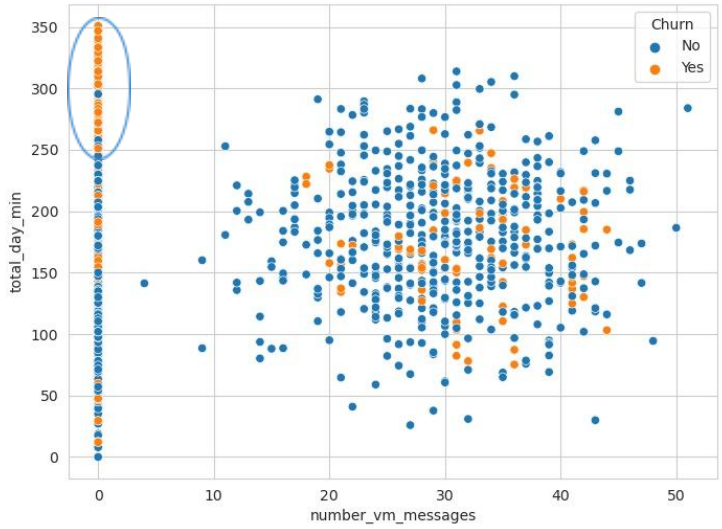


When only considering the customers who have international_plan, you can clearly see using the above picture that a customer who talked international calls for more than 13 minutes or more minutes has churned.

Insight 4: There is a high probability for a customer to churn when he/she talks calls for more minutes in the daytime.



We can come to this insight with the help of the picture below. Among the customers who talked for more than 300 minutes in the daytime, 40 out of 48 were churned customers.



Customers who haven't voice_mail_plan have a high chance to churn when they talk for more minutes in the daytime. Among the customers who talked for more than 300 minutes in the daytime without the voice_mail_plan, 40 out of 43 were churned customers.

The issue might be,

1. You may have been providing poor coverage during the daytime for various reasons, like heavy network traffic during the daytime.
2. You may have been charging more money per minute in the daytime.
3. You may have been charging more money per minute from those who talked for a long time in the daytime.

Insight 5: Within 2312, 575 customers have churned. Within that 575, 488 and 410 customers hadn't voice_mail_plan and intertiol_plan respectively.