MCS - 7103

KAFUUMA HENRY

2024/HD05/21926U 240072196

EXPLORATORY DATA ANALYSIS OF CUSTOMER CHURN IN TELECOM DATASET

ABOUT CUSTOMER CHURN

The high cost of customer acquisition has made telecom operators encounter the “ceiling”, and even fall into the dilemma of customer acquisition. As market saturation increases, telecom operators must solve the problem of increasing subscriber stickiness and prolonging subscriber life cycle. Therefore, it is crucial to analyze and predict the churn of telecom users. Using the data provided, We analyze the data to determine what variables are correlated with customer churn, if any. To identify the customers that might churn.

Research Questions Before the Dataset.

1. What percentage of customers churn?
2. Why do customers churn?
3. How can we prevent churn?
4. What are the control variables?

Research Questions after the Dataset

1. What are the main control variables that determine churn?
2. What variables do not affect customer churn?
3. What variable should be maximized/minimized to reduce churn?

Data Wrangling

In this phase, I analyzed the dataset to determine the data structure i.e., features, type of variables, dimensions, missing values, and basic descriptive statistics.

The data set consists of 21 features and 3333 records. out of which 3 objective, integer, and float both have 9 and 1 Boolean data type. There are no null values or missing data.

29]:

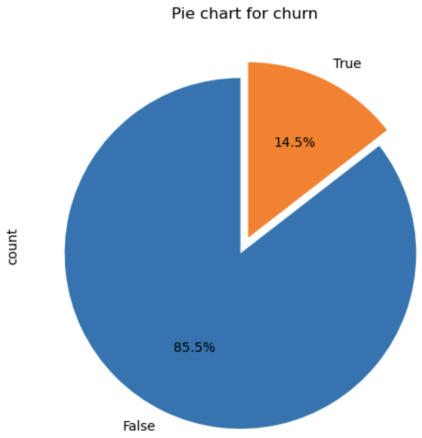
	state	account length	area code	phone number	international plan	voice mail plan	number vmail messages	total day minutes	total day calls	total day charge	...	total eve calls	total eve charge	total night minutes	total night calls	total night charge	total intl minutes	total intl calls	total intl charge	customer service calls	churn
0	KS	128	415	382-4657	no	yes	25	265.1	110	45.07	...	99	16.78	244.7	91	11.01	10.0	3	2.70	1	False
1	OH	107	415	371-7191	no	yes	26	161.6	123	27.47	...	103	16.62	254.4	103	11.45	13.7	3	3.70	1	False
2	NJ	137	415	358-1921	no	no	0	243.4	114	41.38	...	110	10.30	162.6	104	7.32	12.2	5	3.29	0	False
3	OH	84	408	375-9999	yes	no	0	299.4	71	50.90	...	88	5.26	196.9	89	8.86	6.6	7	1.78	2	False
4	OK	75	415	330-6626	yes	no	0	166.7	113	28.34	...	122	12.61	186.9	121	8.41	10.1	3	2.73	3	False

5 rows × 21 columns

EXPLORATORY DATA ANALYSIS (EDA)

Churn column

This is the dependent variable, visualizing using the pie chart we can see the total percentage of churn rate is almost 15%.



ANALYSING DEPENDENT FEATURES vs DEPENDENT VARIABLE

Sate column

Compulsion of state by churn rate - There is 51 unique states present that have different churn rates. From the analysis CA, NJ, TX, MD, SC, and MI are states that have a higher churn rate of more than 21.

Area code column.

There are only 3 unique values.e., 408,415,510, and the churn rate of these area codes is almost the same. There is no kind of relation present between the "area code" and"churn" due to which the customer leaves the operator.

International Plan column

3010 customers don't have an international plan, and 323 customers have an international plan. Among those who have a global plan 42.4 % of people churn. Whereas among those who don't have an international plan only 11.4 % of people churn. So basically customers who bought International plans are churning more than those without it.

Voice Mail Plan Column

From the analysis, there is no clear relation between voice mail plan and churn.

Customer service calls column

From the analysis, customers who called the service center 5 times or above, their churn percentage is higher than 60%. Customers who have called once also have a high churn rate indicating their issue was not solved in the first attempt.

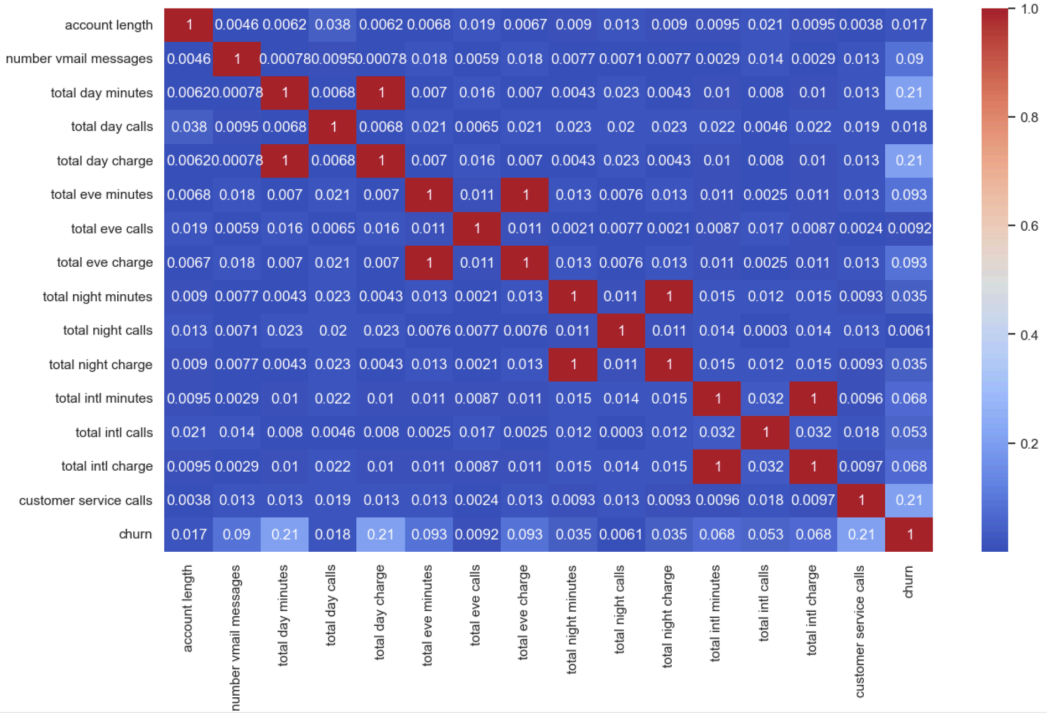
Columns for minutes, calls, and charges

Analyzing all calls, minutes, and charges together, all these are numerical datasets.

After analyzing the dataset we noticed that total day/night/eve minutes/call/charges do not affect the churn rate. But international call charges are high as compare to others it's an obvious thing but that may be a cause for international plan customers to churn out.

MULTIVARIATE ANALYSIS

From correlation matrix using a heatmap to help me understand the correlation between the target variable and each of the other control variables. Total day minutes, total day charge, international minutes, international calls, and customer service features have a slightly higher correlation with customer churn



CONCLUSION:

1. After performing exploratory data analysis on the data set, this is what we have incurred from the data:
2. There are some states where the churn rate is high as compared to others may be due to low network coverage.
3. Area code and Account length do not play any kind of role in the churn rate so, it's redundant data columns
4. In the International plan, customers who have this plan churn more, the international calling charges are also high so the customers who have the plan are unsatisfied with the network.
5. In the voice mail section when there are more than 20 voice-mail messages then there is a churn so it means that the quality of voice mail is not good.
6. Total day call minutes, total day calls, Total day charge, Total eve minutes, Total eve calls, Total eve charge, Total night minutes, Total night calls, and Total night charge, these features didn't play any kind of role regarding the churn rate.
7. In international calls data shows that the churn rate of those customers is high, those who take the international plan churn more, so it means that in international call charges are high.
8. Customer service calls data shows us that the more an unsatisfied customer calls the service center the churn rate is high, which means the service center didn't resolve the customer issue.

RECOMMENDATION

1. Improve network coverage churned state
2. In the international plan provide some discount plan to the customer
3. Improve the voicemail quality or take feedback from the customer
4. Improve the service of the call center take frequent feedback from the customer regarding their issue and try to solve them as soon as possible