SyriaTel Customer Churn Prediction Using Machine Learning and Computer Vision.

Business Understanding.

The telecommunication company is losing customers every month and with everyone wanting to know why there seem to be losing customers they created a Dataset with the people that left and stayed looking for factors that have contributed to their departure.
They also need to know the at risk of leaving groups in order to put some effort into trying to keep them as customers.

Research Question.

The main goal of the project was to use Data Analysis to identify the factors that lead to churning.

Objectives

- Perform Exploratory Data Analysis of the dataset to understand each variable, and their relationship among each other, and to the target the churn
- To identify the variables affecting churn.
- To create a model that relates churn with variables
- Predict the values
- Evaluating our model on how well these it can predict churn .
- Come up with conclusions and recommendations.

Data Understanding.

Data Source.

The Dataset was sourced from kagle but was downloaded and opened locally

Data Description.

The Data had 21 rows with different variables and each had 3333 columns of non-null values and with different Data types the Rows were:

state - the state they are in

account length - How long they have/had the account

area code Area code used for the phone number

phone number The Primary key and unique phone numbers

international plan If they had an international plan

voice mail plan If they had a voicemail plan

number vmail messages The number of voicemails left

total day minutes	Number of minutes they spent calling
total day calls	Total number of calls during the day calls
total day charge	How much they were charged for the daytime calls
total eve minutes	Number of minutes they spent calling in the evening
total eve calls	Total number of calls during the day calls in the evening
total day calls	Total number of calls during the day calls in the day
total eve charge	How much they were charged for the evening time calls
total night minutes	Number of minutes they spent calling in the night
total night calls	the total number of calls during the day calls in the night
total night charge	How much they were charged for the evening time calls
total intl minutes	how many minutes they spent on international call
total intl calls	Total number of intl calls
total intl charge	Total cost for intl calls
customer service calls	Number of times call center were called
churn	If the customer left the telecom

Data Preparation.

Loading the data.

The dataset was stored in a CSV file and loaded into Python using the pandas library with the first 5 rows shown below.

	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	cu
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	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	
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	
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	
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	

5 rows × 21 columns

Cleaning the Dataset

The Dataset was first checked for null values of which none were found

The Dataset was then checked for duplicates of which it had none

The Dataset types for each column were checked and no abnormalities were found.

EDA

Some of the rows could be put together like the charges, minutes, and number of calls which were then binned and run in a plot against churn. Alot of count plots were used to get for example the number of accounts that churned and how many stayed. The data was aso dissected per state and the percentages per state calculated.

Modeling.

The Data was split into test and Train with a Random_va;ue of 1 and the test percentage was left to Default meaning 75% would be used to train and the training 25% would be sed to test/validate. The target variable churn was set to y and the rest of the dataset was set to X

We were able to do a logistic regression on the Dataset since the Target variable was binary(meaning only 2 options were available true or false) and came up with an accuracy score which was a little lower than expected which meant it was under fitted therefore meaning using pipeline which then made it higher but too high meaning it overfit. The third model we used with adding a few hyperparameters is to reduce the accuracy score to a more acceptable place making it the best fit.

CONCLUSIONS

- The customers charged above 40 are at very high risk of churning
- The customers who have higher customer care calls above 4 calls are very likely to churn.
- the highest churn percentage (26.4%) was noticed in CA against the national aVG of (AVG_NAT)
- Account length doesnt reduce churn rate.

Recomendations/further research

- Research why CA,tx and MD percentage for churn is that high (for example CA as to why the churn is 24.X% while the average national is 14%)
- Market and promote international plans as most of the people who receive international calls have no international plan
- Get more information on your customers like Age, gender to understand more demographics of your customers
- Customers with more than 3 customer care calls should have access to managers or people of authority.